July 2024

Stanford | Cyber Policy Center
Freeman Spogli Institute and Stanford Law School

Regulating under Uncertainty: Governance Options for Generative AI

Florence G'sell

NOTA BENE: This study is presented here in draft form. Page and footnote numbers are not final.

ACKNOWLEDGMENTS

This work would not exist without the dedicated efforts of the members of the Governance of Emerging Technologies program at the Cyber Policy Center, directed by **Florence G'sell** and managed by **Ben Rosenthal**. We are greatly indebted to the **Project Liberty Institute** for their support of the Program on Governance of Emerging Technologies, which made this report possible.

This report has benefited from the significant inputs of several key contributors. **Elliot Stewart** focused on technology, meticulously examining the practices of major AI companies, a task made even more challenging by their constantly evolving practices and policies. **Chris Suhler** and **Ashok Ayyar**, assisted by **Nikta Shahbaz**, scrutinized the ongoing strategy of the U.S. federal administration as well as the regulations adopted and proposed at both the federal and state levels. Professor **Jiaying Jiang, Jasmine Shao**, and **Sabina Nong** joined their efforts to provide a comprehensive and precise overview of the Chinese framework. **Zeke Gillman** analyzed the frameworks of Canada, South Korea, Singapore, the UK, Israel, Saudi Arabia, and the Emirates, and also studied ongoing international initiatives. **Arpit Gupta** examined the legal framework of India and the collective practices of AI companies. **Tally Smitas** analyzed the legislation currently being adopted in Brazil. **Ryoko Matsumoto** provided an overview of Japan's framework. Professor **Keeheon Lee** offered significant insights into South Korea's regulatory strategy, while **Nathan Levit** and **Maya Rodrigues** finalized the presentation of the South Korean and Singaporean frameworks.

It is also appropriate to express gratitude to **Sanna Ali**, **Tamian Derivry**, and **Luca Lefevre** for their research efforts and assistance on various topics throughout the drafting of this report.

This report has also been enriched by the valuable feedback and comments of numerous colleagues who dedicated their time to review it and offer substantial and relevant suggestions. In particular, Professor **Jingwen Wang** and Professor **Xinyu Fu** provided invaluable insights on generative AI technology, while **Dave Willner** shared his expertise on the generative AI industry.

This work has also benefited from the precious comments of **Nate Persily, Carlos Escapa, Sarah Cen, David Shao, Jiankun Ni, Sijia Liu, Mick Li, Chun Yu Hong, Hiroki Habuka, Shayne Longpre, Rishi Bommasani, Kevin Klyman, Dan Ho, Mark Lemley, Daphne Keller, Presley Warner, Suzanne Marton, Jerrold Soh, Conor Chapman** and **Samidh Chakrabarti**.

Invaluable technical assistance was provided by copy editors **Lisa Keen** and **Eden Beck**, and designer **Michi Turner**.

Nick Amador, Zeke Gillman, Nathan Levit, Nate Low, Jasmine Shao, Nikta Shahbaz, and **Harith Khawaja** completed the Bluebooking.

Finally, all the students of the Spring 2023 **Governance and Regulation of Emerging Technologies Policy Practicum** should be thanked for their inspiring work, which served as the starting point for this policy report: **Taylor Applegate, Sindy Braun, Alexis Dye, Drew Edwards, Mary Rose Fetter, Dakota Foster, Christopher Giles, August Gweon, Caroline Hunsicker, Poramin Insom, Crys Jain, Harith Khawaja, Atsushi Kono, Ashley Denise Leon, Zahavah Levine, Miranda Lin Li, Katherine McCreery, Caroline Meinhardt, David Mollenkamp, Ilari Papa Kate Reinmuth, Gabriela A. Romero, Greg D. Schwartz, Sade Snowden- Akintunde, Elliot Stahr, Silva Stewart, Christine Strauss, Chris Suhler, Kiran Wattamwar,** and **Ashton E. Woods**..

EXECUTIVE SUMMARY

The revolution underway in the development of artificial intelligence promises to transform the economy and all social systems. It is difficult to think of an area of life that will not be affected in some way by AI, if the claims of the most ardent of AI cheerleaders prove true. Although innovation in AI has occurred for many decades, the two years since the release of ChatGPT have been marked by an exponential rise in development and attention to the technology. Unsurprisingly, governmental policy and regulation has lagged behind the fast pace of technological development. Nevertheless, a wealth of laws, both proposed and enacted, have emerged around the world. The purpose of this report is to canvas and analyze the existing array of proposals for governance of generative AI.

Just as development of the technology has accelerated over the last two years, so too has the debate concerning the relative risks and benefits of generative AI. The "techno-optimists" who celebrate a future of pervasive AI point to exponential increases in human productivity, discovery, and creativity. Detractors worry about everything from severe economic and environmental impacts to existential risks from autonomous AI agents and new weapons of mass destruction. This very unsettled debate over the relative risks and benefits of AI casts a cloud of uncertainty over the policy debate related to regulation. As the technology emerges from its infancy, the rules considered and enacted now, with incomplete information, could have significant impacts on the trajectory of technological development.

Although much of the policy debate, as expected, focuses on risks, most governments recognize the important benefits AI promises for their economies and population. Although no consensus has yet emerged as to the transformational potential of the current crop of generative AI, few dispute its widespread economic and social impact. Indeed, dramatic scientific breakthroughs have already been realized. For example, DeepMind's development of AlphaFold to predict protein structures has made dramatic progress in the field of computational biology. These early successes have led the most optimistic AI supporters to predict that AI will make possible any number of drug and medical breakthroughs that could cure or treat endemic diseases. They see AI providing technological solutions to seemingly intractable global challenges, such as climate change, pollution, and energy shortages. But the benefits will not be limited to scientific discovery. All sectors of the economy should expect some impact from AI —from data-intensive businesses to all creative industries. And as AI becomes "embodied" with robotics, the potential shall be realized in all forms of manual labor as well.

Of course, some of these benefits are also viewed as risks or costs. AI's transformation of the economy, like all previous technological transformations, will come with massive job displacement. And while AI may help find ways to mitigate climate change and energy shortages, the building of massive data centers and the training of new models promises to significantly increase energy demands in the short term due to the buildout of AI technology. Those who warn of the societal impacts from AI are alarmed by the potential for catastrophic harms —from novel viruses and new weapons to uncontrollable AI agents and cyberattacks. Early in the immediate aftermath of ChatGPT, leading AI scientists and business leaders called on

governments to begin addressing collectively the problems of new existential risks posed by AI. However, experts remain divided on the plausibility of "loss of control" scenarios, where a highly intelligent "rogue AI" could surpass human oversight and potentially spiral out of control. And critics of that concern over future existential risks suggest focus should be placed instead on the immediate and tangible risks posed by generative AI that require government action. Disinformation threats from synthetic imagery, an explosion in virtual child pornography, scams using voice mimicking technology to defraud unsuspecting victims, discrimination and bias in AI algorithms, and a host of problems due to vulnerabilities to jailbreaking and inaccurate responses provided by chatbots are just a few of the problems already presented by existing generative AI tools. As these tools are rolled out for use in law enforcement, criminal justice, judicial process, employment, education, healthcare, and any number of other domains, both the malfunctioning of the systems and their abuse by bad actors cautions against overreliance on AI and wholesale replacement of human oversight of these processes.

Governance of generative AI, therefore, requires policymakers to walk the difficult line between enabling the tremendous benefits of AI while warding off both present day harms and future existential risks. Proposed and existing government regulation occurs along a continuum, from a laissez faire model, that mostly characterizes the United States, to a more command-and-control model characteristic of traditional forms of regulation, with China at the extreme opposite pole from the U.S. In the middle are different degrees of co-regulation, such as that prevalent in the European Union, in which governments exist in a dialogic relationship with companies to respond incrementally to new developments and discovered harms from the technology.

To be clear, though, government regulation of some kind is inescapable, if only to clarify how existing laws will apply to the newest technologies. One of the most pressing issues relates to the fact that developers train their models by using extensive datasets, often gathered through online web scraping, a practice which may often scoop up copyrighted content or personal data. Multiple lawsuits have been filed around the world, alleging infringement by generative AI models —trained on copyrighted text and images— which sometimes produce AI-generated content that is remarkably similar to that of existing creative works. The law will have to determine whether the training of AI models brings with it the risk of intellectual property infringement. It will have to ascertain if a generative AI user, by merely writing a prompt, can be declared the author of whatever the AI tool produces in response to that prompt. It will also have to elucidate the implications of current practices by AI companies concerning privacy and data protection. The law of libel and defamation will need to adapt to the AI environment. When a model produces defamatory content about an individual, who should be held liable? Given that chatbots cannot produce foolproof answers to information requests, what steps (if any) should be taken to immunize companies from reputational and other harms created by use of such tools?

Therefore, although the United States may be characterized as having a "hands-off" regulatory approach, either through the courts or through legislation, some regulation will be necessary. For now, most of the action in the US has occurred either in the executive branch or in the states. In addition to securing voluntary commitments from the major AI companies, the Biden administration issued Executive Order 14110, which outlines eight guiding principles and policy priorities for federal agencies and authorities. The

Executive Order instructs various federal departments to issue reports, develop guidelines, and take actions under existing authority, as well as improve AI oversight in the federal government. Consequently, the National Institute of Standards and Technology (NIST) has published various documents detailing non-binding standards on AI risks and cybersecurity, developed through collaboration with the AI industry and stakeholders. The absence of a comprehensive legal framework has prompted some individual states to enact their own AI-related legislation, addressing issues such as deepfakes and algorithmic discrimination. Perhaps the most significant proposal, California's pending "Safe and Secure Innovation for Frontier Artificial Intelligence Models Act," would impose significant compliance obligations on AI developers.

The most significant and comprehensive piece of legislation passed thus far with respect to AI is the EU's AI Act. This legislation follows in a long line of recent European technology regulation: privacy protection through the General Data Protection Regulation (GDPR); digital platform content regulation through the Digital Services Act (DSA); competition regulation through the Digital Markets Act (DMA); and intellectual property regulation through the Copyright Directive. Some provisions in those laws, to a greater or lesser degree, apply to AI. In addition, some EU Member States, similar to some states in the United States, have their own laws that may apply to AI. Moreover, the AI Act is not the only European legislation adopted recently that implicates AI. The Cyber Resilience Act imposes several cybersecurity rules on "products with digital elements," and the new Product Liability Directive classifies software, including AI systems, as products that can be defective, thereby holding their producers strictly liable in case of defectiveness.

The AI Act regulates AI systems based on risk levels and use cases, particularly in sensitive sectors. The Act categorizes risks based on the "intended" use of AI systems and classifies them into four risk categories: unacceptable risk, high risk, limited or "transparency" risk, and minimal or no risk. Depending on the risk level involved in the application, different legal requirements will apply.

- Unacceptable –
    - Examples: social scoring, person-based crime prediction
    - Requirements: Such AI systems are prohibited
- High risk –
    - Examples: critical infrastructure, education, health care and medicine, product safety, employment, law enforcement, border control, administration of justice and democracy.
    - Requirements: risk management; data quality and governance, consistent recordkeeping, transparency and provision of information to deployers, guarantee of human oversight, ensuring system accuracy, robustness, and cybersecurity
- Limited Risk –
    - Examples: chatbots, generative AI systems
    - Requirements: transparency to ensure that humans are informed when they are interacting with AI and labeling of AI-generated content.

- Minimal or No Risk –

- Examples: video games and spam filters.
- Requirements: No specific regulation of AI.

During the negotiation process over the AI Act, provisions were added to regulate general-purpose (i.e., foundation) models, shifting the focus from specific use cases to the technology itself. Consequently, these general-purpose AI models, like GPT, Llama, or Claude, have numerous applications and are trained on massive amounts of data. Developers of these models must provide technical documentation, including information about the model's energy consumption, documentation for downstream providers, greater transparency relating to their training data sets, as well as a copyright policy. General-purpose AI models posing systemic risks must comply with additional obligations related to cybersecurity, red teaming, risk mitigation, incident reporting, and model evaluation.

This report spends a considerable amount of space on the European Union's AI Act because it presents many of the fundamental regulatory choices policymakers must confront. Parts of the law represent a command-and-control model, specifically mandating rules for the development, deployment, and application of AI. The Act also has aspects of co-regulation to it. For example, stakeholders are involved in enforcement by developing codes of practice. The AI Act grapples with the challenges of regulating the technology itself, as well as the applications of AI. It also has certain exceptions for open-source models. Like other European tech regulations, the impact of the EU AI Act will be felt beyond Europe as it applies to all AI services used in the EU, which, as with GDPR, may set a new international standard for AI. Finally, the AI Act is also important because several other countries, such as Brazil, are emulating it.

Aggressive state regulation of AI, as with any technology, can come in different forms. China presents one paradigm for authoritarian regimes. It has enacted an array of laws that place substantial constraints on the development of the technology. These include the *Algorithm Recommendation Provisions*, the *Deep Synthesis Regulation*, the *Interim Measures for Generative AI Services*, and the *Basic Safety Requirements for Generative AI Services*. Some of the goals in these laws are similar to those in democratic regimes. For instance, they call for labeling of AI-generated imagery and protection of privacy and intellectual property, at least with respect to the development of the technology by non-state actors and its impact on other citizens. However, for China, the laws are specifically crafted to ensure that the technology furthers national values. Reporting obligations for covered technologies (which extend well beyond typical generative AI applications to include all kinds of algorithms) include reporting to regulators on technology and uses that are *capable* of influencing public opinion or mobilizing the public. Users are subject to close monitoring, while generative AI tools must be trained and configured to minimize the production of "fake news" and "illegal or unhealthy information." Service providers must promptly address any infringements and implement measures to prevent future occurrences, such as by "optimizing the training" of AI models.

Given the geopolitical implications of the race to control this new and powerful technology and the ease with which the technology will transcend the boundaries of a given regulator, it should come as no surprise that most major international organizations have proposed or are drafting new initiatives related to AI. The G7's *Hiroshima AI Process* has resulted in non-binding yet influential frameworks, such as its *Guiding Principles* and *Code of Conduct* for AI. The G20 has also published its own *G20 AI Principles*, and the EU-US Trade and Technology Council has released collaborative AI projects. The five-nation BRICS group

has formed an AI Study Group to foster innovation and establish AI governance standards, in alignment with China's *Global AI Governance Initiative*. In May 2024, the Council of Europe adopted the first international AI treaty, the *Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law*, which requires ratification by at least five states to enter into force. The United Nations established a High-Level Advisory Body on AI and adopted its first AI resolution. UNESCO provided ethical guidelines and global guidance on AI use in education and research. Additionally, the African Union published various policy documents to guide AI development in Africa. Most of these international initiatives have been strongly influenced by the work of the OECD's "Recommendation on Artificial Intelligence."

Several high-level principles and observations emerge from this exploration of the different national and international initiatives related to AI:

1. **Regulation of the technology or its applications.** Sector specific laws allow AI regulation to work incrementally and adapt laws in narrow ways to the changes made by AI. However, due to the uncertain reach and implications of AI technology and the development of general-purpose AI models, predicting future applications and use cases is challenging. Consequently, regulating the technology itself is particularly important. This necessity explains why some countries, such as the EU, have adopted specific provisions targeting AI models. Additionally, many countries, including the United Kingdom, the United States, Japan, and the European Union, have established dedicated AI Safety Institutes.

2. **The importance of transparency and auditing**. Because the impacts of generative AI are unknown, transparency in the development of this technology is critical. "Model cards" and disclosures about training data represent only the beginning of the necessary transparency. To fully understand the implications of foundation models and generative AI applications, both developers and external third parties must rigorously test them prior to deployment. This testing should aim to evaluate performance, biases, alignment, and the potential to generate significant risks.

3. **Regulations and enforcement.** Given the complexity and rapid pace of development of the technology, legislation can go only so far in specifying rules *ex ante* that will govern AI development and applications, even in the near future. Enforcement will prove as important, if not more so, than legislation. This will require governments to hire AI talent, which is both expensive and in short supply. It will also require government coordination with companies and civil society to provide continuous guidance on how the rules on the books apply to new and emerging contexts.

4. **The relative power of the public and private sectors.** Almost all current generative AI models have been developed by private companies. The need to collect vast amounts of data, the scarcity of chips, and the high costs of computation have concentrated the resources required to develop and train the most powerful models in the hands of a few private companies.

To "democratize" the production of AI may require massive public investment to ensure that actors other than those tied to the profit-maximizing mission of the commercial firms will be able produce the most cutting edge AI models.

5. **The promise and risks of open models.** Although private companies are developing the most powerful models, some are publicly releasing the models and their weights. Meta and its Llama models have taken a lead in the production of powerful open-source models, although others, such as Mistral or Falcon, have released significant open-source models. Open-source models promise to make the benefits of AI available to the world. They also might create a competitive environment, quite different from that of social media and search engines, which have been controlled by a few oligopolistic actors. However, the openness of these models also generates significant concerns. Once the models are released, they can be used and fine-tuned by bad actors for all kinds of intended and unintended purposes. Moreover, once they are "out the door," there is little that the companies or governments can do to control their impact. Government regulation in this space must address the relative risks and benefits posed by open-source models.

The title of this report – "Regulating Under Uncertainty: Governance Options for Generative AI" – seeks to convey the unprecedented position of governments as they confront the regulatory challenges AI poses. Regulation is both urgently needed and unpredictable. It also may be counterproductive, if not done well. However, governments cannot wait until they have perfect and complete information before they act, because doing so may be too late to ensure that the trajectory of technological development does not lead to existential or unacceptable risks. The goal of this report is to present all of the options that are "on the table" now with the hope that all stakeholders can begin to establish best practices through aggressive information sharing. The risks and benefits of AI will be felt across the entire world. It is critical that the different proposals emerging are assembled in one place so that policy proponents can learn from one another and move ahead in a cooperative fashion.

Table of Content

Chapter 1 Introduction *13*

1.1. Regulating under uncertainty **13**

1.2. Possible regulatory approaches **19**

1.2.1. Self-regulation 19

1.2.2. Co-regulation 22

1.2.3. Traditional government regulation 24

1.3. Positions within the industry **28**

1.3.1. Calls for regulation 28

1.3.2. Calls for international initiatives 32

1.4. Purpose and structure **33**

Chapter 2 Generative AI: The technology and supply chain *35*

2.1. What is generative AI? **35**

2.1.1. Artificial Intelligence 35

2.1.2. Generative Artificial Intelligence 37

2.2. Developing generative AI models **46**

2.2.1. Machine-Learning techniques 46

2.2.2. Model pre-training 47

2.2.3. Fine-tuning 49

2.2.4. Model architecture 51

2.2.5. Resources required for development 52

2.3 The supply chain **55**

2.3.1. Upstream providers vs. downstream deployers and users 55

2.3.2. Open-Source vs. Closed-Source release 58

2.3.3. Profitability models 63

KEY TAKEAWAYS **64**

Chapter 3 Challenges and risks of generative AI *67*

3.1. Technical and operational risks **67**

3.1.1. Technical vulnerabilities 68

3.1.2. Factually incorrect content 71

3.1.3. Opacity 77

3.2. Ethical and social risks **81**

3.2.1. Malicious use and abuse 81

3.2.2. Misinformation and disinformation 88

3.2.3 Bias and discrimination 89

3.2.4. Influence, overreliance, and dependence 92

3.2.5 Nascent capabilities 96

3.2.6. Risk disparities among different models 106

3.3. Legal challenges **111**

3.3.1. Privacy and data protection concerns 111
3.3.2. Copyright challenges 114

3.4. Environmental, economical, and societal challenges 119
3.4.1. Concentration of market power 119
3.4.2. Impact on labor markets 124
3.4.3. Environmental cost 128
3.4.4. Artificial General Intelligence 131

KEY TAKEAWAYS 136

Chapter 4 Industry initiatives 139

4.1. Industry practices 139
4.1.1. Pre-deployment safety practices 140
4.1.2. Deployment safety practices 161
4.1.3. Post-deployment safety practices 166

4.2. Collective initiatives 185
4.2.1. The Partnership on AI 185
4.2.2. Frontier Model Forum 188
4.2.3. The AI Alliance 189
4.2.4. MLCommons 189
4.2.5 Coalition for Content Provenance and Authenticity 190
4.2.6. Other initiatives 191

KEY TAKEAWAYS 192

Chapter 5 Regulatory initiatives 197

5.1. The European Union 197
5.1.1. Existing legal frameworks in the EU 198
5.1.2 The AI Act 224
5.1.3. The liability directives 276
5.1.4. The Cyber Resilience Act 293

KEY TAKEAWAYS 296

5.2. China 300
5.2.1. General overview of China's AI strategy 300
5.2.2. Data and copyright protection 302
5.2.3 China's regulatory initiatives on AI technologies 307
5.2.4 China's main initiatives on AI ethics 327
KEY TAKEAWAYS 331

5.3. The United States 337
5.3.1 Existing legal frameworks 337
5.3.2. US federal regulatory initiatives 358
5.3.3. State regulatory initiatives in the US 389

KEY TAKEAWAYS 398

5.4. Ongoing regulatory initiatives 405

5.4.1. Brazil 405
5.4.2. Canada 410
5.4.3. India 418
5.4.4 Israel 427
5.4.5. Japan 429
5.4.6. Kingdom of Saudi Arabia 435
5.4.7. Singapore 437
5.4.8. South Korea 443
5.4.9. The United Arab Emirates 449
5.4.10. United Kingdom 451

KEY TAKEAWAYS 461

Chapter 6 International initiatives and negotiations 463

6.1. United Nations 464
6.1.1. High-Level AI Advisory Body 464
6.1.2. General Assembly AI resolutions 465
6.1.3. Global Digital Compact 465

6.2. OECD 466
6.2.1. The OECD's recommendation on Artificial Intelligence 467
6.2.2. OECD's publications 469
6.2.3. The AI Policy Observatory 470

6.3. The G7 471
6.3.1. Content of the *Hiroshima AI Process Comprehensive Policy Framework* 472
6.3.2. Impact of the *Hiroshima AI Process Comprehensive Policy Framework* 474
6.3.3. Other recent developments 475

6.4. The G20 476
6.4.1. The 2019 Osaka summit and the release of the *G20 AI Principles* (2019) 476
6.4.2. Subsequent G20 summits 477

6.5. BRICS 478

6.6. African Union 479

6.7. AI safety summits 481
6.7.1. The UK AI Safety Summit (November 2023) 481
6.7.2. The AI Seoul Summit 484

6.8. The Council of Europe's treaty 485
6.8.1. Drafting of the Council of Europe's AI treaty 485
6.8.2. Key features of the treaty 487
6.8.3. Implementation of the treaty 489
6.8.4. Limitations 490

6.9. The Global Partnership on AI 491

6.10. US-EU Trade and Technology Council 492
6.10.1. The *Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management* 493

6.10.2. Privacy-enhancing technologies 494
6.10.3. AI's impact on the workforce 495
6.10.4. The AI Code of Conduct 496
6.11. UNESCO 496
6.11.1. Recommendation on the ethics of AI (2021) 496
6.11.2. Guidance on generative AI in education and research 498
KEY TAKEAWAYS 499
Chapter 7 Final conclusions 502

Chapter 1 Introduction

Effective regulation of emerging technologies inevitably presents legislators with a set of difficult choices. If they act aggressively to mitigate all hypothetical risks, they might inhibit the development of the technology. If they act too conservatively at the outset, they might miss the chance to steer the industry toward the safe development of the technology and away from foreseeable harms. These choices must also be made at a time when the knowledge and expertise about the technology resides mostly in the private sector. As a result, governments often do not have the necessary expertise to design and enforce a new regulatory regime.

Perhaps more than any previous technology, artificial intelligence illustrates these regulatory trade-offs and challenges. Institutions inside and outside of government, including the AI companies themselves, are clamoring for legislation. Creating "rules of the road" is seen as necessary both for the prevention of harm and for the establishment of ground rules of the growing market for the technology. Consensus may exist around abstract principles, as seen in various voluntary commitments or guidelines, but that consensus often breaks down once discussions turn to implementation with concrete regulations that seek to strike the right balance.

This report seeks to describe the different ways that legislators, regulators, and non-governmental actors have grappled with the problem of regulating generative AI. It presents a variety of proposals and approaches at varying levels of specificity in an attempt to assess the regulatory landscape as of June 2024. It begins with a description of the technology and the risks that regulation might seek to prevent. It then presents industry-led efforts at self-regulation, before turning to a discussion of proposals from around the world, with a particular focus on the European Union's AI Act, the most comprehensive legislative initiative passed thus far. The report then discusses the proposals from various international bodies before presenting a series of conclusions.

1.1. Regulating under uncertainty

Predictions of the future promised by artificial intelligence range from the apocalyptic to the utopian. There are those who see in the technology risks at the scale of the Terminator, with the threat of human extinction at the hands of killer robots, bioweapons, or a host of other intended or unintended innovations. Then, there are those who see it as humanity's saving grace, with the potential to solve a range of human problems, from cancer to climate change. As the capabilities of this technology rapidly become apparent, legislators are placed in the almost-impossible position of trying to preserve the upsides of this emerging technology while avoiding the potentially catastrophic consequences. This is all the more difficult because of the considerable uncertainty, at this early stage, as to when and how the most significant dangers and benefits will become readily apparent. To some extent, governments are legislating in the dark, as they set the ground

rules for a technology presenting hypothesized existential risks, some present-day dangers, and potential long-term socially transformative benefits.

1.1.A. The ChatGPT turning point

Historians may look back on November 30, 2022, as the critical turning point in the development of generative AI. Although beforehand, the technology was in widespread use, particularly in academia, it was on that date that OpenAI released ChatGPT to the world. When it did so, it shot the starting gun for what has turned into an "AI arms race,"[1] in which technology companies now compete to develop the most sophisticated models as quickly as possible.[2] In March 2023, four months following the release of ChatGPT in November 2022, OpenAI launched GPT-4, a model with dramatically enhanced capabilities. Concurrently, Anthropic introduced Claude, a direct competitor to GPT-4.[3] Meta unveiled its open model, Llama 2, in July 2023, Mistral AI released Mistral 7B in September 2023, Baidu launched Ernie 4.0 one month later, followed by X.ai's release of Grok in November and Google's release of Gemini at the close of 2023. The momentum continued into 2024, with Anthropic releasing Claude 3 in March, Meta introducing Llama 3 in April, and OpenAI releasing GPT-4o in May.

At present, only a select few AI companies, often global tech giants, possess the expertise and financial resources necessary to develop the most advanced generative AI systems and models (*see section 2.2.5.*). Having invested in AI for many years, these leading AI companies are well-positioned to dominate the burgeoning generative AI market. Yet small- and medium-sized enterprises (SMEs), startups, researchers, and open-source developers are also playing a role in the generative AI ecosystem. In particular, while most leading AI companies opt to keep the proprietary details of their AI systems private, an increasing number of developers are choosing to be open.[4] The presence of these open models facilitates a potentially competitive market that distinguishes this technological ecosystem from the monopolistic tendencies of recent technological advancements, for example, in social media and search engines. At the same time, the openness of the models necessarily presents unique risks, as the model developers cannot foresee or rein in the efforts of bad actors who might use these models to create harm (*see section 3.2.6.A.*).

Competition in AI is not limited to companies, of course: It extends to nation states. Because of the potential power of this new technology, let alone its military and intelligence applications, the AI arms race has similarities to an actual arms race, as countries compete to develop offensive and defensive capabilities to promote national interests. The AI revolution, therefore, has far-reaching geopolitical implications. In the

---

[1] Kevin Roose, *How ChatGPT Kicked Off an A.I. Arms Race*, N.Y. TIMES (Feb. 3, 2023), https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html.

[2] In February 2023, Microsoft CEO Satya Nadella said, "Rapid innovation is going to come. In fact, a race starts today." Richard Waters & Madhumita Murgia, *Microsoft targets Google's search dominance with AI-powered Bing*, FINANCIAL TIMES (Feb. 8, 2023), https://www.ft.com/content/2d48d982-80b2-49f3-8a83-f5afef98e8eb.

[3] The notable releases of 2023 are presented in Stanford's *Artificial Intelligence Index Report*, STAN. U. HUMAN-CENTERED ARTIFICIAL INTELLIGENCE (2024), at 78–80, https://aiindex.stanford.edu/report/ [hereinafter Stanford AI Index Report 2024].

[4] "Since 2011, the number of AI-related GitHub projects has seen a consistent increase, growing from 845 in 2011 to approximately 1.8 million in 2023. Notably, there was a sharp 59.3% rise in the total number of GitHub AI projects in the last year alone." *Id.* at 69.

words of a Brookings Institution 2020 blog post: "Whoever leads in artificial intelligence in 2030 will rule the world until 2100."[5]

The competition in AI is now often seen as a two-way race between the leading technological and economic superpowers, the United States and China. They are the clear front-runners in AI development, even though Europe also holds a significant position and a few other countries, such as the United Arab Emirates, have made significant investments. According to the Stanford AI Index Report 2024, 61 notable AI models were developed by US-based institutions in 2023, while 21 were developed in the European Union and 15 in China.[6] The concentration of advanced AI model development in only a few developed countries raises concerns about this "Global AI Divide"[7] that could render other countries dependent on critical technologies.

1.1.B. Benefits

Over the past two decades, AI has progressively transformed various fields, showcasing significant advancements and benefits. One notable milestone was AlphaGo, developed by DeepMind, which defeated the world champion Go player in 2016. This achievement demonstrated the potential of AI in mastering complex tasks through deep learning and reinforcement learning techniques, marking a significant leap in AI capabilities. Another groundbreaking development was in the field of protein folding. In 2020, DeepMind's AlphaFold achieved a major breakthrough by predicting protein structures with remarkable accuracy, solving a 50-year-old challenge in biology (*see below section 3.2.1.C.*).

Concerning generative AI in particular, its capabilities have rapidly expanded beyond mere content creation. Generative AI models are now adept at a wide array of tasks and can serve as the foundation for developing various specialized AI systems. With rapidly advancing capabilities, they are able to accomplish increasingly sophisticated tasks. One example, among many, was again provided by DeepMind: In December 2023, its AI model solved a previously unsolved mathematical problem. In a paper published in *Nature*,[8] researchers stated that the accomplishment marked the first instance of a large language model being used to solve a long-standing unsolved mathematical problem. The solution was not present in the training data and was entirely novel.[9]

---

[5] Indermit Gill, *Whoever leads in artificial intelligence in 2030 will rule the world until 2100*, BROOKINGS (Jan. 17, 2020), https://www.brookings.edu/articles/whoever-leads-in-artificial-intelligence-in-2030-will-rule-the-world-until-2100.

[6] Stanford AI Index Report 2024 *supra* note 3 at 47.

[7] Yoshua Bengio et al., *International Scientific Report on the Safety of Advanced AI, Interim Report* (May 2024), at 57, https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf [hereinafter *International Scientific Report*].

[8] Bernardino Romera-Paredes et al., *Mathematical discoveries from program search with large language models*, NATURE 625, 468–75 (2024), https://doi-org.stanford.idm.oclc.org/10.1038/s41586-023-06924-6.

[9] Will Douglas Heaven, *Google DeepMind used a large language model to solve an unsolved math problem*, MIT TECH. REV. (Dec. 14, 2023), https://www.technologyreview.com/2023/12/14/1085318/google-deepmind-large-language-model-solve-unsolvable-math-problem-cap-set/.

Current generative AI tools can now produce various types of output, e.g.: text, images, computer code, music, videos, or even structure synthesis for 3D printing. A novel activity known as "prompt engineering" has emerged, involving the optimization of textual input to enhance communication with a generative AI tool, though future AI systems may render this unnecessary by becoming more intuitive and proficient in understanding natural language.[10] In the creative industries, generative AI enhances content creation by generating images, music, and text, allowing artists and writers to experiment with new styles and ideas. It plays a crucial role in education by creating tailored learning experiences and materials, adapting to the needs of individual students. Businesses may benefit from generative AI through improved customer service, as it powers advanced chatbots and virtual assistants capable of handling complex inquiries.

In the future, generative AI systems will be integrated into myriad products and services and applied in areas such as customer support, artistic creation, image enhancement, research initiatives, coding, and virtual assistants for driving. Generative AI offers the potential to revolutionize multiple sectors—education, entertainment, healthcare, and scientific research—by enabling the creation of customized, scalable content, automating processes, generating hypotheses, and boosting efficiency. In software development, generative AI accelerates the coding process by generating code snippets and analyzing existing code to suggest optimizations. It also enhances data analysis, producing synthetic data that can be used to train models when real data are scarce or sensitive. Generative AI supports scientific research by generating hypotheses and simulating experiments, saving time and resources. In healthcare, it can assist in drug discovery and the development of personalized treatment plans by simulating complex biological processes. Finally, generative AI can aid in the detection and prevention of fraud by generating realistic scenarios that help identify vulnerabilities in security systems.

Some characterize the opportunities offered by generative AI as extraordinary, with the expectation that the technology will lead to significant scientific breakthroughs, economic growth, and profound social transformations. The most extreme pronouncement of the utopian vision of AI comes from Marc Andreessen's Techno-Optimist Manifesto: "We believe technology is liberatory. Liberatory of human potential. Liberatory of the human soul, the human spirit. Expanding what it can mean to be free, to be fulfilled, to be alive. We believe technology opens the space of what it can mean to be human."[11] In July 2023, the British Computer Society (BCS) appeared to agree with this approving assessment. The BCS released an open letter calling for the UK government and industry to recognize AI as "a transformational force for good, not an existential threat to humanity."[12] The letter attracted over 1,300 signatures.

---

[10] "Prompt engineering is the process of writing, refining and optimizing inputs to encourage generative AI systems to create specific, high-quality outputs." IBM, *What is prompt engineering?*, THINK (2024), https://www.ibm.com/topics/prompt-engineering. Among many studies, *see* Sander Schulhoff et al., T*he Prompt Report: A Systematic Survey of Prompting Techniques*, arXiv (June 6, 2024), https://arxiv.org/abs/2406.06608. *See also* Oguz A. Acar, *AI Prompt Engineering Isn't the Future*, HARVARD BUSINESS REVIEW (June 6, 2023), https://hbr.org/2023/06/ai-prompt-engineering-isnt-the-future.

[11] Marc Andreessen, *The Techno-Optimist Manifesto*, ANDREESSEN HOROWITZ (Oct. 16, 2023), https://a16z.com/the-techno-optimist-manifesto/.

[12] British Computer Society, *BCS Open letter calls for AI to be recognised as 'force for good not threat to humanity,'* BCS (July 18, 2023), https://www.bcs.org/articles-opinion-and-research/bcs-open-letter-calls-for-ai-to-be-recognised-as-force-for-good-not-threat-to-humanity/.

1.1.C. Fears and worries: A pause

Since the release of ChatGPT, the excitement surrounding generative AI has been modulated by concerns the technology —open or closed source— has raised. In OpenAI CEO Sam Altman's words, "if this technology goes wrong, it can go quite wrong."[13] Many worry that AI companies, in their rush to innovate, may overlook the importance of ensuring their models are safe and aligned with human values. This anxiety led numerous researchers and industry leaders, such as Elon Musk and Steve Wozniak, to sign an open letter[14] in March 2023, calling on "all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4." Endorsed by over 33,000 signatories, the letter urged that stakeholders slow down, if not stop, the technological race. One signatory said the letter was provoked, at least in part, by an alleged "corporate irresponsibility,"[15] presumably referring to a lack of self-regulation by major AI companies.

Importantly, the open letter argued that companies like OpenAI are "locked in an out-of-control-race to develop and deploy ever more powerful digital minds that no one —not even their creators— can understand, predict, or reliably control." The proposed pause, the signers wrote, should allow independent experts and AI companies to generate a set of shared industry safety protocols and address various issues. At a minimum, said the letter, these should include:

- instituting oversight and tracking of highly capable AI systems and large pools of computational capability,
- implementing provenance and watermarking mechanisms to differentiate synthetic media from authentic media and to track model leaks,
- establishing an auditing and certification framework,
- devising liability systems for AI-caused harm,
- allocating government funds for technical AI safety research, and
- organizing new proficient regulatory bodies focused on artificial intelligence.

"Powerful AI systems should be developed," the letter highlighted, "only once we are confident that their effects will be positive and their risks will be manageable."

In the following months, a succession of other experts issued public statements, some even more dramatic than the open letter. In May 2023, numerous AI experts endorsed a statement by the Center for AI Safety, emphasizing that "mitigating the risk of extinction from AI should be a global priority alongside other

[13] Ryan Tracy, *ChatGPT's Sam Altman Warns Congress That AI 'Can Go Quite Wrong,'* WALL ST. J. (May 16, 2023), https://www.wsj.com/articles/chatgpts-sam-altman-faces-senate-panel-examining-artificial-intelligence-4bb6942a.

[14] PAUSE GIANT AI EXPERIMENTS: AN OPEN LETTER (Mar. 22, 2023), https://futureoflife.org/open-letter/pause-giant-ai-experiments/; Cade Metz & Gregory Schmidt, *Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society,'* N.Y. TIMES (Mar. 29, 2023), https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.

[15] Metz & Schmidt, *supra* note 14.

societal-scale risks such as pandemics and nuclear war."[16] Additionally, the Executive Committee of the Conference on Fairness, Accountability, and Transparency (FAccT) and over 250 FAccT community members signed a Statement on AI Harms and Policy in June 2023: "Our research has long anticipated harmful impacts of AI systems of all levels of complexity and capability….This moment calls for sound policy based on the years of research that has focused on this topic."[17]

To date, however, there has not been any significant move to pause the release of cutting-edge advanced models or to curb the expansion of AI capabilities. Many, inside and outside the industry, have emphasized the impracticality of the pause advocated by the open letter.[18] OpenAI CEO Sam Altman did acknowledge that AI companies should have "increasing rigor for safety issues," but he faulted the open letter for lacking "technical nuance."[19] Notably, Meta recently launched Llama 3, Anthropic introduced Claude 3, and the release of OpenAI's GPT-5 is imminent. Rather than halting progress, the concerns expressed by the calls for a pause served to highlight the necessity for private companies to self-monitor and mitigate risks associated with developing increasingly powerful AI systems.

1.1.D. Striking the right balance: Innovation vs. harm-prevention

Concerns about AI safety are as old as AI; they have persisted since machine learning significantly augmented the capabilities of AI models. Recent years have seen an array of concerns emerge, including the potential misuse of tools with advanced capabilities, the inherent biases of AI tools trained on low-quality datasets, and the lack of full explainability in these systems.

Generative AI introduces additional challenges. The main feature of generative AI tools is their ability to produce high-quality, original content, which poses new issues regarding the nature and potential impact of the newly created content. Additionally, the broad scope and versatility of generative AI models bring significant uncertainties concerning the range of possible applications. The capability to fine-tune these models allows various downstream developers and users to adjust them in ways that could have severe and harmful outcomes. This is particularly concerning as increasingly powerful models are being released to

---

[16] Center for AI Safety, *Statement on AI Risk* (May 2023), https://www.safe.ai/work/statement-on-ai-risk#open-letter.

[17] ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), *Statement on AI Harms and Policy*, FAccT, https://facctconference.org/2023/harm-policy (last visited June 20, 2024).

[18] *See* Herb Scribner, *Key lines from Elon Musk, others' call to pause AI development*, Axios (Mar. 29, 2023), https://www.axios.com/2023/03/29/elon-musk-gpt-4-chat-open-ai; *see also* Jennifer Rigby, *Bill Gates says calls to pause AI won't 'solve challenges,'* Reuters (Apr. 4, 2023), https://www.reuters.com/technology/bill-gates-says-calls-pause-ai-wont-solve-challenges-2023-04-04/; DeepLearningAI, *Yann LeCun and Andrew Ng: Why the 6-month AI Pause is a Bad Idea*, YouTube (Apr. 7, 2023), https://www.youtube.com/watch?v=BY9KV8uCtj4; On the other hand, notable AI ethicists argued that, by ignoring the present harms of already deployed AI systems, the letter did not go far enough. Devin Coldeway, *Ethicists fire back at 'AI Pause' letter they say 'ignores the actual harms,'* TechCrunch (Mar. 31, 2023), https://techcrunch.com/2023/03/31/ethicists-fire-back-at-ai-pause-letter-they-say-ignores-the-actual-harms/.

[19] Rohan Goswami, *OpenAI Sam Altman Addresses Letter from Musk and Other Tech Leaders Calling for A.I. Pause*, CNBC (Apr. 14, 2023), https://www.cnbc.com/2023/04/14/openai-ceo-altman-addresses-letter-from-musk-wozniak-calling-for-ai-pause.html.

the market, and advanced models may develop unforeseen and potentially dangerous capabilities (*see section 3.2.5.B.*). Furthermore, generative AI poses a number of risks . These risks include, but are not limited to, the propagation of disinformation, the creation of deepfakes, and the production of other manipulated content, all of which can have serious consequences (*see chapter 3*).

The challenge for policymakers lies in achieving an optimal equilibrium between fostering innovation for the collective good and ensuring robust safety and risk mitigation measures. Furthermore, their policies must exhibit sufficient agility to keep abreast of the rapid advancements in a context where AI models are being introduced at an accelerating rate.

1.2. Possible regulatory approaches

For decision-makers, particularly policymakers, determining the most effective strategy for addressing an emerging innovation becomes challenging when the technology's potential risks and future trajectory are still unclear. Of course, since AI is deployed on a global scale, the instinctive response is to consider the implementation of a comprehensive global AI governance strategy. Such an approach necessitates international cooperation, which can sometimes lead to significant initiatives. For example, the G7 adopted a code of conduct (*see section 6.3.*), and international AI safety summits have produced important statements, such as the Bletchley Declaration (*see section 6.6.1.B.*). In practice, however, AI governance policies primarily develop at the state level or within highly structured supranational organizations, such as the European Union or the African Union.

Within this framework, one potential strategy for governments is a laissez-faire approach, which entails allowing innovation to progress naturally in the expectation that market forces will automatically promote economic growth and risk reduction. Alternatively, a more interventionist approach could compel private entities to adopt specific rules and standards aimed at mitigating current or anticipated risks associated with the technology. The most decisive option available to policymakers is to implement binding regulations without delay, thereby establishing clear guidelines for the development of the technology.

1.2.1. Self-regulation

For policymakers, adopting a laissez-faire approach entails relying on the market and private sector to foster the optimal development of an innovation. This strategy presumes that technology companies are the most qualified to foresee and evaluate risks and that they have a vested interest in addressing these risks effectively. The freedom granted to the industry under such an approach may have positive effects. In the case of generative AI, the lack of regulation has enabled a swift pace of innovation, evidenced by the rapid succession of model releases over the past 18 months. For the public, this translated into quick access to cutting-edge technologies, benefiting end users and enabling businesses to enhance efficiency and competitiveness.

Meanwhile, it is somewhat common for companies to share information and collaborate to pinpoint relevant technical standards and best practices. This culture of exchange and collaboration is notably active within the technology sector and among AI model developers. Collaboration among companies can result in enhanced cooperation and the collective creation of best practices that progressively become standard across the entire industry. This forms the basis for "self-regulation" initiatives, although the term encompasses a range of practices and realities that can differ significantly.

1.2.1.A. What is self-regulation?

According to the Organisation for Economic Co-operation and Development (OECD), industry self-regulation "concerns groups of firms in a particular industry or entire industry sectors that agree to act in prescribed ways, according to a set of rules or principles."[20] These industry groups may develop best practice frameworks for risk governance, monitor adherence, and even enforce compliance. Company participation in these groups is generally voluntary. The standards issued by industry groups may have varying degrees of legal force: They may be purely voluntary or formally binding.

For instance, in July 2023, Anthropic, Google, Microsoft, and OpenAI announced the formation of the Frontier Model Forum,[21] a new industry body dedicated to the safe and responsible development of advanced AI models (*see section 4.2.2.*) The core objectives of the Forum include advancing AI safety research; identifying best practices; and collaborating with policymakers, academics, civil society, and companies to share knowledge about trust and safety risks. However, the work of the Forum does not result in any binding standards.

The industry can also develop codes of conduct to which companies voluntarily commit. These codes of conduct may be drafted by industry representatives and experts from professional organizations. For example, the U.S. National Academy of Medicine is collaborating with leading organizations in health, bioethics, equity, technology, patient advocacy, and research to develop an Artificial Intelligence Code of Conduct (AICC). This initiative aims to outline the national framework necessary to foster and support the equitable and responsible use of AI in health, medical care, and health research.[22]

Additionally, companies may, collectively or individually, make voluntary commitments on specific matters. In July and September 2023, leading US AI companies pledged to adhere to several best practices, including watermarking, in response to a request from the White House (*see section 5.3.2.B.2.*). While it might be argued that, in this case, these commitments transcend self-regulation because they involve the

---

[20] Organisation for Economic Co-operation and Development (OECD), *Industry Self-Regulation: Role and Use in Supporting Consumer Interests*, DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INNOVATION COMMITTEE ON CONSUMER POLICY (Mar. 23, 2015), at 11, https://one.oecd.org/document/DSTI/CP(2014)4/FINAL/En/pdf .

[21] Google, *Frontier Model Forum: A new partnership to promote responsible AI*, THE KEYWORD (July 26, 2023), https://blog.google/outreach-initiatives/public-policy/google-microsoft-openai-anthropic-frontier-model-forum/; *see also* George Hammond, *Top Tech Companies Form Group Seeking to Control AI*, FIN. TIMES (July 26, 2023), https://www.ft.com/content/709f4375-83bf-4037-878d-964d1ead8858.

[22] National Academy of Medicine, *NAM Leadership Consortium Collaborates with Leading Health, Tech, Research, and Bioethics Organizations to Develop Health Care AI Code of Conduct* (June 20, 2023), https://nam.edu/nam-leadership-consortium-collaborates-with-leading-health-tech-research-and-bioethics-organizations-to-develop-health-care-ai-code-of-conduct/.

government, they remain within the realm of self-regulation due to their nonbinding nature. Certainly, they could be characterized as a form of "encouraged self-regulation."

1.2.1.B. Advantages of self-regulation

Self-regulation presents advantages.[23] Specifically, it ensures that those with the deepest knowledge of the technology and practical field experience are the ones establishing the standards and best practices for effective and pragmatic risk governance. Industry developers have a greater degree of expertise and technical knowledge of practices than governments. Furthermore, these industry players have the capacity to adapt to the rapid pace of AI development, as they continuously monitor and implement technological innovations on a daily basis. Another argument in favor of self-regulation is that the cost of enforcing such standards is borne by the industry itself, in one way or another, rather than by the taxpayers who fund the regulatory agencies. Ultimately, the standards developed by the industry may provide a valuable foundation for governments when they establish a regulatory framework, as these standards originate from entities with expert knowledge in the field.

1.2.1.C. Limitations of self-regulation

The challenge of self-regulation lies in ensuring that industry standards and company commitments are genuinely upheld —a task that is challenging without a dedicated independent body. While independent watchdogs and nongovernmental organizations (NGOs) can monitor corporate behavior, their efforts alone may be insufficient. However, forming industry-specific bodies, such as professional organizations, can help ensure companies effectively implement these guidelines. Additionally, establishing certification mechanisms, such as those confirming that an AI model was developed in compliance with specific standards for dataset curation and testing, is also a viable strategy.

Nevertheless, relying on the industry may seem excessively optimistic, considering that the primary objective of industry players is to generate profit and capture market share. In the AI sector, companies operate in a highly competitive environment, characterized by a race to develop increasingly sophisticated and powerful AI models. Consequently, they are necessarily tempted to prioritize enhancing performance, often at the expense of mitigating risks and fostering responsibility. In such a context, self-regulation standards and practices may be driven by self-interest. Due to the relative lack of external constraints, self-regulation maximizes opportunities for rent-seeking, which involves obtaining wealth transfers without contributing to productivity or creating new wealth.[24]

At a minimum, an independent authority must be established to inspect industry practices and confirm compliance with established standards. Without such oversight, it is infeasible to guarantee that companies fulfill the commitments they have made or adhere to the codes of conduct they have agreed to. Self-regulatory frameworks lacking an authoritative body to enforce the application of standards and best

[23] A. Ogus, *Rethinking Self-Regulation*, *in* A READER ON REGULATION 174–88 (Robert Baldwin et al. ed., 1998).
[24] *Id.*; *see also* Alyssa Wong, *Regulatory gaps and democratic oversight: On AI and self-regulation*, SCHWARTZ REISMAN INST. FOR TECH. & SOC'Y (Sept. 21, 2023), https://srinstitute.utoronto.ca/news/tech-self-regulation-democratic-oversight.

practices are likely to result in scenarios where AI companies, while professing to adopt "ethical" or "responsible" policies, are actually engaging in strategies solely aimed at enhancing their public image.

Currently, the AI industry remains relatively autonomous in crafting and implementing practices to identify and address the risks associated with its own products. Unfortunately, the exceptionally swift advancement of generative AI and its rapidly increasing availability to the public have led to problematic side effects before effective mitigations could be implemented. Among them are the creation of virtual child sexual abuse material[25] and the spread of offensive content.[26] Meanwhile, major AI companies have become more and more secretive about their development processes, driven by genuine safety concerns, the need to protect their source codes and datasets from competitors, and the desire to protect themselves from potential liability claims. This opacity complicates efforts by independent watchdogs or the public to evaluate the associated risks or to verify whether AI companies are adequately addressing these risks (*see section 3.1.3.B.*).[27]

1.2.2. Co-regulation

A more stringent approach than voluntary self-regulation entails regulators and state agencies playing an active role in developing and effectively implementing standards and best practices. This does not necessarily mean that self-regulation is replaced with traditional regulation. It is common for self-regulatory schemes to include some level of government involvement, which can vary greatly between different frameworks. But the government's implication in self-regulation may extend to "co-regulation," which represents a midpoint in the continuum between self-regulation and full government regulation.

---

[25] An investigation showed that popular AI image generation models were trained with images of child sexual abuse material present in a public dataset of billions of images, known as LAION-5B. Christoph Schuhmann et al., *LAION-5B: An open large-scale dataset for training next generation image-text models*, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (Oct. 16, 2022), https://arxiv.org/pdf/2210.08402. Models trained on this dataset were used to create photorealistic AI-generated nude images, including CSAM. David Thiel, *Investigation Finds AI Image Generation Models Trained on Child Abuse*, STAN. CYBER POLICY CENTER (Dec. 20, 2023), https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

[26] For example, in February 2024, Google's tool, Gemini, produced historically inaccurate images, such as depictions of America's Founding Fathers as Black, the Pope as a woman, and a Nazi-era German soldier with dark skin. The problem was apparently due to a "fine-tuning error." Bobby Allyn, *Google CEO Pichai says Gemini's AI image results "offended our users,"* NPR (Feb. 28, 2024), https://www.npr.org/2024/02/28/1234532775/google-gemini-offended-users-images-race.

[27] Cliff Saran, *Self-Regulation of AI is Not an Option*, COMPUTERWEEKLY (Nov. 25, 2021), https://www.computerweekly.com/blog/Cliff-Sarans-Enterprise-blog/Self-regulation-of-AI-is-not-an-option ("Among the areas of concern is that unlike traditional research, which is steeped heavily in academia, half of the research papers on AI are coming out of commercial research outfits. This is a double-edged sword. On the one hand, commercial research is driving adoption of advanced AI in business. However, unlike academic research, there is a risk that being commercially sensitive, the source code and datasets used in these AI algorithms and models cannot easily be reviewed independently."); *see also* Rishi Bommasani *et al.*, *Improving Transparency in AI Language Models: A Holistic Evaluation*, STAN. U. HUMAN-CENTERED AI (Feb. 2023), https://hai.stanford.edu/sites/default/files/2023-02/HAI%20Policy%20%26%20Society%20Issue%20Brief%20-%20Improving%20Transparency%20in%20AI%20Language%20Models.pdf ("Language models developed and used by companies like Google and Microsoft in search engines, content moderation, and translation services may be closed–meaning they are not accessible to regulators and external researchers, limiting outsiders' ability to understand the system.").

Co-regulation can take various forms. While participation in industry groups, such as the Partnership on AI or the Frontier Model Forum (*see below section 4.2.*), is typically voluntary, it could be mandated by law. Another option is for the government or an independent public authority to approve rules issued by private entities or industry groups, thereby making the rules mandatory. For instance, codes of conduct may be developed jointly with industry players and other stakeholders. In May 2016, the EU Commission collaborated with Facebook, Microsoft, Twitter, and YouTube to establish a "Code of Conduct on Countering Illegal Hate Speech Online" to prevent and combat the spread of illegal hate speech on the internet.[28] Its implementation is evaluated through a regular monitoring exercise conducted by various organizations across different EU countries.[29]

Another example of co-regulation comes from the recently adopted EU's AI Act framework.[30] The AI Office, established within the EU Commission in January 2024, is responsible for encouraging the development of codes of practice to implement the obligations of providers of general-purpose AI models (*see section 5.1.2.C.*). These codes are to be drafted by the industry in collaboration with relevant national authorities, civil society organizations, industry, academia, and other stakeholders, including downstream providers and independent experts. Providers of general-purpose AI models will be invited to adhere to these codes of practice and may use them to demonstrate compliance with the AI Act. EU authorities will monitor effective compliance with the codes.[31] The AI Act integrates traditional regulatory mechanisms with co-regulation techniques (*see section 5.1.2.E.*).

In the US, in July 2023, the U.S. National Institute of Standards and Technology (NIST)[32] established the "Generative AI Public Working Group" to spearhead the development of a cross-sectoral AI Risk Management Framework (RMF) profile for managing the risks associated with generative AI models and systems.[33] The efforts of this working group, which comprised over 2,500 members, have informed the creation of the draft AI RMF Generative AI Profile released on April 29, 2024, to address the risks associated with the specific use case of generative AI (*see section 5.3.2.B.3.c.ii.*).[34] Meanwhile, NIST has recently established the Artificial Intelligence Safety Institute Consortium (AISIC) for bringing together AI

---

[28] European Commission, *The EU Code of conduct on countering illegal hate speech online*, https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (last visited June 13, 2024).

[29] *Id.*

[30] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). PE/24/2024/REV/1. OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj, (*see below section 5.1.2.*)

[31] The AI Office and the European Artificial Intelligence Board (*see below section 5.1.2.E.1.*).

[32] National Institute of Standards and Technology, *About NIST*, NIST, https://www.nist.gov/about-nist (last visited Apr. 1, 2024) ("NIST is a division of the U.S. Department of Commerce whose mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology.").

[33] National Institute of Standards and Technology, *NIST AI Public Working Groups*, TRUSTWORTHY & RESPONSIBLE AI RESOURCE CENTER, https://airc.nist.gov/generative_ai_wg (last visited June 13, 2024).

[34] National Institute of Standards and Technology, *AI Risk Management Framework*, INFORMATION TECHNOLOGY LABORATORY, https://www.nist.gov/itl/ai-risk-management-framework (last visited June 13, 2024).

developers and users, academics, government and industry researchers, and civil society organizations.[35] The Consortium already unites over 200 organizations and will develop guidelines and standards for AI measurement and policy. The guidelines published within these frameworks are not legally binding; however, companies that choose to adhere to them are strongly encouraged to comply. Although there is no formal mechanism to enforce compliance, the expectation is that adherence will be undertaken with a high degree of commitment, since these standards were developed in collaboration with a federal agency.

Effective collaboration between industry and regulators in the development of standards can bring significant benefits. It can combine industry's technological expertise with regulators' commitment to user protection and legal compliance. The participation of regulators in the development of predominantly industry-driven governance practices ensures that the objectives of protecting the general public and ensuring safety are considered. This cooperation can also enable the creation of processes to oversee industry practices and enforce standards. For instance, an independent organization or regulatory agency may be tasked with monitoring the implementation of codes of conduct. Utilizing a commonly agreed-upon methodology, these organizations can assess how well the companies are fulfilling their commitments.

However, these co-regulatory mechanisms do not totally exclude the risk of regulatory capture, where the industry, possessing deeper technical expertise —particularly in an environment of swift technological advancement— can influence the establishment of standards that primarily serve its own interests. Moreover, most of these co-regulatory mechanisms fall short of producing effectively mandatory rules or methods for ensuring that AI companies genuinely adhere to their commitments and announced policies.

1.2.3. Traditional government regulation

The last option for policymakers is to opt for regulation in its traditional form: enacting a law or some form of binding legal framework. However, when dealing with emerging technologies, this decision is far from straightforward.[36] It is challenging for lawmakers to certainly determine what legal framework will be most appropriate without knowing whether the benefits will outweigh the potential dangers.

1.2.3.A. Objections to government regulation

The need for regulation that aligns with the current state of the technology and its challenges may be hampered by the potential lack of expertise among legislators, who frequently lack essential knowledge about technological advancements. Governments must seek expertise to acquire the essential information required for drafting well-adapted regulations. However, in the field of AI, expertise predominantly resides within private companies, which, rather than academic institutions, are developing the most advanced models. According to Stanford's AI Index, academia was at the forefront of releasing machine-learning

[35] National Institute of Standards and Technology, *Artificial Intelligence Safety Institute Consortium (AISIC)*, https://www.nist.gov/artificial-intelligence-safety-institute/artificial-intelligence-safety-institute-consortium-aisic (last visited June 13, 2024).

[36] Lyria Bennett Moses, *How to Think About Law, Regulation and Technology: Problems with 'Technology' as a Regulatory Target*, 5(1) LAW, INNOVATION AND TECHNOLOGY 1–20 (2013), https://ssrn.com/abstract=2464750.

models until 2014, after which the industry took the leading role.[37] For instance, in 2023, the industry produced 51 notable machine-learning models, compared to only 15 from academia.[38] As a result, the discussion on AI regulation is taking place in a context where the relevant expertise lies mostly in the hands of strictly private and profit-oriented entities. This information asymmetry between regulators and AI companies creates a risk of regulatory capture. If regulators get biased information from industry experts, they might draft laws and regulations that serve the industry's private interests.

Moreover, even without regulatory capture, crafting balanced and adaptable legal frameworks is particularly challenging when technology evolves rapidly. This difficulty is exacerbated when the capabilities and potential impacts of the technology are not well understood. A lack of clear understanding of the potential effectiveness of such regulations in addressing specific risks may result in poorly targeted regulatory burdens.[39] Specifically, laws that are either misdirected or overly stringent can significantly stifle innovation. In the context of AI, stringent regulatory requirements, such as those related to training data or model testing, could impose substantial costs on developers. If the burden of regulations leads to significant compliance costs for the regulated businesses, it is likely that only large, established companies with substantial revenues will be able to bear these expenses. In contrast, smaller and less wealthy companies, unable to manage such high compliance costs, may struggle to enter the market or find it much more challenging to compete.[40]

Furthermore, even if those who draft the law possess a thorough understanding of the technological landscape and market conditions at the time of regulation, their law may soon become obsolete. For instance, the current provisions of the GDPR, adopted in 2016, appear to be ill-suited to the characteristics of generative AI (*see section 5.1.1.A.*). Consequently, the challenge of applying this law effectively and pragmatically generates significant legal uncertainty. Such circumstances are detrimental not only to users but also to the whole economy, which relies on cutting-edge technology for development.

Ultimately, beyond the drafting of laws, the need for expertise is critical during the implementation and enforcement phases. Regulatory agencies must possess robust technical knowledge to guarantee effective application of the established rules. Without experts who can monitor technological advancements in real time, the effectiveness of enforcement may be limited. One potential solution is close cooperation between the industry's leading companies and regulatory agencies, but this approach raises, once more, the risk that AI companies do not disclose reliable information or withhold important details. An alternative approach is to integrate independent experts into the enforcement mechanism. For instance, the recently enacted EU Digital Services Act (DSA)[41] introduces a unique monitoring system to enforce compliance with its

---

[37] Stanford AI Index Report 2024 *supra* note 3 at 46.

[38] While 21 significant models emerged from industry-academic collaborations, the 15 most noteworthy models were all developed by the industry. *Id.* at 78–80.

[39] Neel Guha et al., *AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing*, GEO. WASH. L. REV. (forthcoming 2024), https://ssrn.com/abstract=4634443.

[40] Some scholars have highlighted this possible detrimental side effect, particularly regarding the GDPR. *See* Mary Fan, *The Hidden Harms of Privacy Penalties*, 56 U.C. DAVIS L. REV. (2022), https://ssrn.com/abstract=4143821.

[41] Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) PE/30, 2022 O.J. (L 277) ELI: http://data.europa.eu/eli/reg/2022/2065/oj.

obligations. Article 40 of the DSA states that very large online platforms and search engines must provide internal data on request to researchers vetted by national regulators. These researchers may request whatever data they need to assess systemic risks and propose mitigation strategies. While it remains uncertain whether the system implemented by the DSA will prove truly effective, a similar mechanism for involving the academia could be considered for AI governance, as advocated by numerous researchers and experts in the field.[42] The idea is to allow independent researchers and auditors to analyze models with full access in a secure environment, thereby facilitating independent assessment efforts.

1.2.3.B. Advantages of government regulation

On the other hand, opting to regulate—for instance, by enacting a law to govern the primary deployment and use of technology—seems to be far more effective in ensuring safety and mitigating harm than relying on self-regulation. The creation of a legislative framework not only allows for the formulation and clarification of binding rules for all players but also enables the establishment of specific enforcement mechanisms. When carefully crafted, these mechanisms can ensure effective oversight of technology companies and guide the trajectory of technological development. Furthermore, the establishment of an authority, such as a regulatory agency dedicated to enforcement, ensures a significant degree of transparency, since such an authority can compel technology companies to disclose information and data they might otherwise withhold. Transparency is essential for assessing and mitigating potential risks.

1.2.3.C. Forms of government regulation

Once the decision to regulate has been made, the next question involves determining the appropriate forms of regulation. This is a particularly delicate issue, as regulating a complex and rapidly evolving technology carries the risk of "regulatory misalignment." A recent paper by Guha et al. highlights this concern, asserting that such misalignment occurs when regulatory goals or unintended consequences fail to address the harms they target or introduce unacknowledged trade-offs between different objectives.[43] The paper analyzes the technical and institutional feasibility of four commonly proposed AI regulatory regimes — disclosure, registration, licensing, and auditing— and concludes that each of these regimes suffers from its own regulatory alignment issues. It suggests that all AI-related concerns (such as mandating transparency, fairness, privacy preservation, accuracy, and explainability) cannot be simultaneously achieved.

In any case, the primary task for the legislator is to identify the goals that the proposed law should achieve. Should it uphold essential principles following a principle-based approach? Or should the drafters of the law take a more pragmatic stance, focusing primarily on risk mitigation with a risk-based approach? The two approaches are not necessarily mutually exclusive. The EU, for instance, has adopted a risk-based strategy while still upholding certain principles, such as human oversight (*see section 5.1.2.*). Conversely,

[42] More than 350 members of AI, legal, and policy communities have called for companies to provide "safe harbor" for good faith research and evaluation activities. *A Safe Harbor for Independent AI Evaluation, Open letter*, MIT, https://sites.mit.edu/ai-safe-harbor/ (last visited June 16, 2024).

[43] "Effective and clear regulation requires clarity about the nature of the harm (or market failure) a regulation is seeking to address." Guha et al., *supra* note 39.

China has adopted a principle-based approach by imposing general principles and rules that are independent of actual risk levels (*see section 5.2.3.*).

The second question that must be addressed by those seeking to regulate an emerging technology is whether the *technology* should be regulated or its *applications*. The focus on *technology* is based on the premise that the technology may be inherently dangerous and that its risks can be managed only through precise regulation of its developers. With *applications*, the assumption is that the technology is not necessarily dangerous and that any potential hazards arise from its use, putting the focus on deployers and users. When a law specifically targets technology, it must provide clear, precise technical definitions and specifications. This task can be challenging, as exemplified by the AI Act's relatively broad and vague definition of "general purpose AI models" (*see section 5.1.2.C.*) Conversely, focusing on applications necessitates that regulators anticipate the possible use cases of the technology and their potential risks —a challenging task given the rapid and sophisticated advancements in emerging technologies. In any case, these two approaches, once again, are not mutually exclusive. The AI Act, for example, incorporates both types of provisions: one set focusing on use cases determined by sectors and classified by their degree of risk, and another set addressing "general purpose AI models," which are considered to present particular risks due to their advanced capabilities (*see below section 5.1.2.*).

Third, the decision to regulate necessitates a careful arbitration between various options for substantive measures. In the case of AI, numerous questions arise. The first question is to determine the process for releasing AI models and systems. Should developers be permitted to release their models without oversight? Should they self-declare and self-assess, as outlined in the AI Act? (*see section 5.1.2.*) Or should they be subject to a prior control system and obtain the green light from a regulatory agency to release their models, as provided by the licensing system proposed by some US lawmakers? (*see section 5.3.2.C.1.b.*) Another important issue concerns the appropriate regulatory regime for open-source models and applications as opposed to closed-source models (*see section 3.2.6.A.*). There is also the question of whether a specific regulatory regime should be established for the most powerful and capable models (*see section 3.2.6.B.*). Another complex issue is determining the measures to foster innovation, such as implementing favorable policies for SMEs or creating mechanisms to alleviate regulatory burdens, such as regulatory sandboxes. The law must also delineate whether liability should reside with infrastructure providers, downstream deployers, or end users.

Lastly, enforcement mechanisms must be precisely defined. Legislators must determine whether to establish a specialized agency with specific expertise in AI or to empower existing authorities to enforce the law. Substantial investment is needed to ensure that the competent agencies have the necessary resources and expertise to oversee the activities and practices of AI companies. The legal penalties must be sufficient to motivate compliance, even among companies with significant financial resources. And the technical implementation of the law must be meticulously planned, as any legal framework concerning technology must be enforced at a technical level. Regulations must thoroughly detail the technical specifications for the audits and safety tests that will be required.

1.3. Positions within the industry

Since the release of ChatGPT, there has been a vigorous debate surrounding the opportunity and feasibility of regulating artificial intelligence, particularly the most advanced AI models. Leaders in the industry have called for an in-depth dialogue between governments and AI companies,[44] in order to encourage state regulation and launch international initiatives.

1.3.1. Calls for regulation

Many AI companies and industry leaders advocate for regulation.[45] In a recent Op-Ed, two former members of OpenAI's board state that "self-governance cannot reliably withstand the pressure of profit incentives."[46] The authors argue that, even with the best intentions, self-regulation will become unenforceable without external oversight, especially when faced with strong profit incentives. "Governments must play an active role," they concluded.

Naturally, industry leaders who support regulation exhibit varying opinions on the nature and type of rules needed. Most agree on the necessity to strike the right balance so that these legal frameworks do not impose an excessive burden or cost on AI companies, especially on the smallest ones.[47] OpenAI's Sam Altman, for instance, has insisted on the necessity for AI companies to collaborate with the government and the public on developing precautionary measures, with particular attention to addressing the dangers of the massive dissemination of fake news.[48] He also urged the US government to establish a comprehensive list of safety standards, require companies to undergo independent audits, and create a new licensing agency to ensure compliance with these standards.[49] Meta's Mark Zuckerberg insisted on the need for regulation to control the most serious risks, notably disinformation and fake imagery, with particular emphasis on rules requiring watermarking.[50]

---

[44] *See* TIME, *Microsoft CEO Satya Nadella on AI*, YOUTUBE (May 10, 2023), https://www.youtube.com/watch?v=ckIs_HRPmUM (3:55 answer to "[I]f you were in the government what would you be doing to ensure there's enough regulation to protect citizens from AI?").

[45] Cecilia Kang, *OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing*, N.Y. TIMES (May 16, 2023), https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

[46] Helen Toner & Tasha McCauley, *AI firms mustn't govern themselves, say ex-members of OpenAI's board*, THE ECONOMIST (May 26, 2024), https://www-economist-com.stanford.idm.oclc.org/by-invitation/2024/05/26/ai-firms-mustnt-govern-themselves-say-ex-members-of-openais-board.

[47] John Thornhill, *AI will never threaten humans, says top Meta scientist*, FINANCIAL TIMES (Oct. 18, 2023), https://www.ft.com/content/30fa44a1-7623-499f-93b0-81e26e22f2a6.

[48] Cristine Criddle & Hannah Murphy, *OpenAI chief says new rules are needed to guard against AI risks*, FINANCIAL TIMES (May 16, 2023), https://www.ft.com/content/aa3598f7-1470-45e4-a296-bd26953c176f.

[49] *Oversight of A.I.: Rules for Artificial Intelligence: Hearing Before the Subcomm. on Privacy, Technology, and the Law. Comm. on the Judiciary*, 118th Cong. (2023) (written testimony of Sam Altman, CEO, OpenAI) , https://www.judiciary.senate.gov/imo/media/doc/2023-05-16%20-%20Bio%20&%20Testimony%20-%20Altman.pdf [hereinafter Sam Altman Testimony].

[50] Johana Bhuiyan, *Tech leaders agree on AI regulation but divided on how in Washington forum*, THE GUARDIAN (Sept. 13, 2023), https://www.theguardian.com/technology/2023/sep/13/tech-leaders-washington-ai-saferty-forum-elon-musk-zuckerberg-pichai; *see also* Jillian Deutsch, *Zuckerberg, Altman Offer Support for EU Regulation of AI*,

In a US Senate hearing in July 2023, Dario Amodei, who left OpenAI to found the AI company Anthropic, expressed deep concern over the potential malicious use of AI and pushed for a "testing and auditing regime" for new and powerfulAI models.[51] New AI models would have to pass "a rigorous battery of safety tests both during development and before being released to the public."[52]

For its part, Google has published "Recommendations for regulating AI," in which it recommends taking a sectoral approach to AI regulation that builds on existing frameworks (i.e., regulation of industries such as healthcare and financial services) by regulating the specific applications of AI, rather than AI itself.[53] Google also suggests adopting a proportionate, risk-based methodology, promoting interoperable AI standards and governance, ensuring parity between AI and non-AI systems, and recognizing transparency as a means to an end.

1.3.1.A. Licensing

A recurring proposition among industry leaders is licensing for larger AI models. The idea was first popularized by Sam Altman during his May 2023 congressional hearing, where he said that the US government should consider "a combination of licensing or registration requirements for development and release of AI models above a crucial threshold of capabilities."[54] Altman proposed developing safety standards through a multistakeholder approach and implementing external validation mechanisms for AI systems that require licenses or registration. He even suggested that policymakers consider implementing licensing regulations on a global scale and ensure international cooperation on AI safety. This would involve examining potential intergovernmental oversight mechanisms and establishing international standards.

The licensing proposal was later echoed by other industry executives, among them Brad Smith, the president of Microsoft,[55] and Mustafa Suleyman, a co-founder of Deepmind.[56] But each of their proposals lack precision on how an AI licensing regime would operate. In general, a licensing regime would entail the creation of a government agency to ensure "responsible and skilled development and use of AI products, either by licensing companies or practitioners, or through approval of the development or deployment of

---

BLOOMBERG (June 23, 2023), https://www.bloomberg.com/news/articles/2023-06-23/meta-is-well-prepared-to-meet-europe-content-rules-breton-says.

[51] *Oversight of A.I.: Principles for Regulation: Hearing Before the Subcomm. on Privacy, Technology, and the Law. Comm. on the Judiciary*, 118th Cong. (2023) (written testimony of Dario Amodei, Ph.D., Co-founder and CEO, Anthropic), https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf.

[52] *Id.*

[53] Google, *Recommendations for Regulating AI*, at 2–3, https://ai.google/static/documents/recommendations-for-regulating-ai.pdf.

[54] Sam Altman Testimony *supra* note 49 at 12.

[55] Steven Overly, *It's time to regulate AI like cars and drugs, top Microsoft exec says*, POLITICO (Sept. 13, 2023), https://www.politico.com/news/2023/09/13/its-time-to-regulate-ai-like-cars-and-drugs-top-microsoft-exec-says-00115445.

[56] Mustafa Suleyman, *Containment for AI*, FOREIGN AFFAIRS (Jan. 23, 2024), https://www.foreignaffairs.com/world/containment-artificial-intelligence-mustafa-suleyman.

systems themselves."[57] The government would develop minimum safety standards for relevant parties to follow in order to achieve a license.

Microsoft's Chief Executive Officer Satya Nadella[58] and President Brad Smith further promoted the idea of AI licensing within Microsoft's *Governing AI: A Blueprint for the Future* report.[59] The *Blueprint* report paints the contours of a potential US licensing regime for highly capable AI models. There, the company states that, should a licensing regime be implemented, Microsoft would share its industry knowledge to support it.[60] Microsoft explains that the government should not only license highly capable AI models but should license AI data centers, which will be used to test and deploy high-risk AI systems (the report specifies that "high-risk" should be defined by the government). The report also recommends licensing those AI systems that autonomously control critical infrastructure (electrical grids, water systems, city traffic flows, etc.).[61] While the report does little to spell out the requirements for licensing, it does note some important considerations. Among them is that a system should include a second and separate layer of protection for ensuring human control in the event that application-level measures (safety breaks) fail.[62]

Members of the US Congress have taken notice. Senator Lindsey Graham (R-SC), along with Senator Elizabeth Warren (D-MA), proposed the creation of an independent regulatory commission with licensing authority over dominant tech platforms, including those developing AI.[63] More recently, Senators Richard Blumenthal (D-CT) and Josh Hawley (R-MO) have as their first consideration in the "Bipartisan Framework for U.S. AI Act" the creation of an AI licensing regime (*see section 5.3.2.C.1.b.*).[64] The draft bill would require companies to register advanced general-purpose AI models or models used in high-risk situations with a new, independent oversight agency. Deploying these models would require a license. To acquire a license, a company would need to meet three basic requirements: disclosing and sharing relevant information on the model, following certain compliance measures (risk management, pre-deployment testing, data governance, and adverse incident reporting mechanisms), and facilitating agency audits.

1.3.1.B. The Microsoft Blueprint: "Know Your Cloud, Customers, and Content"

Beyond the licensing principle, Microsoft's *Blueprint* report provides a global proposal for an AI regulatory framework. The proposals in the *Blueprint* are based on setting up a regulatory architecture that is similar

---

[57] Guha et al., *supra* note 39.
[58] Tracy, *supra* note 13.
[59] Microsoft, GOVERNING AI: A BLUEPRINT FOR THE FUTURE at 19-21 (May 25, 2023), https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw.
[60] *Id.* at 20.
[61] *Id.* at 14.
[62] *Id.* at 14–15.
[63] *Warren, Graham Unveil Bipartisan Bill to Rein in Big Tech*, WARREN.SENATE.GOV (JULY 27, 2023), https://www.warren.senate.gov/newsroom/press-releases/warren-graham-unveil-bipartisan-bill-to-rein-in-big-tech.
[64] Senators Richard Blumenthal and Josh Hawley, *Bipartisan Framework for U.S. AI Act* (Sept. 7, 2023) , https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf [hereinafter *Bipartisan Framework for U.S. AI Act*].

to the architecture of AI systems, so that the implemented framework is adapted to the sector.[65] It also implies the implementation of the KY3C principle—a strategy adapted from the banking industry: "Know Your Customer (KYC)." Microsoft's proposal explains that, under existing law, financial institutions are required to verify customer identities, establish risk profiles, and monitor transactions to help detect possible illegal activity. The comparable principle for the AI industry, says the *Blueprint*, could be "Know Your Cloud, Customers, and Content," or "KY3C."[66] This would ensure that regulatory principles can be effectively followed at all levels of the production chain.

First and foremost, the *Blueprint* proposes that the government define risk levels and critical infrastructure. This first level of action is necessary because it determines the risk factors and the actors who should be subject to a high level of regulation.[67] Secondly, Microsoft suggests that regulations should focus on ensuring that developers of AI systems at risk or destined for use in critical infrastructure include safety brakes in the designs of their models.[68] These safety brakes, modeled on aviation or nuclear practices, should enable alerts to be issued and system action to be stopped immediately when a loss of control appears imminent. Thirdly, Microsoft suggests that high-risk AI systems, or those destined for use in critical infrastructure, should undergo a thorough testing phase before being put into service, under a government-controlled procedure.[69]

In this *Blueprint*, Microsoft is positioning itself as open to regulation and as a bridge builder between government and industry.[70] This is a noteworthy strategy on the part of a company with a particularly strong presence in the world of AI, especially given its close partnership with OpenAI.[71]

1.3.1.C. Paradoxical stances

Although the leading figures in the AI industry have expressed support for regulation, certain paradoxes remain. In May 2023, OpenAI's Altman hinted at the possibility of withdrawing OpenAI products from the European Union because of the EU's challenging regulatory environment.[72] Later, he reversed this stance, confirming that OpenAI had no plans to exit the EU market and speaking positively about ongoing discussions on AI regulation in Europe.[73] Although Mark Zuckerberg previously emphasized the necessity

[65] *Id.* at 17–18.
[66] *Id.* at 6.
[67] *Id.* at 14.
[68] *Id.*
[69] *Bipartisan Framework for U.S. AI Act supra* note 64 at 14.
[70] Cat Zakrzewski, *Microsoft won over Washington. A new AI debate tests its president*, THE WASHINGTON POST (May 25, 2023), https://www.washingtonpost.com/technology/2023/05/25/brad-smith-microsoft-ai/.
[71] Microsoft Corporate Blogs, *Microsoft and OpenAI extend partnership*, OFFICIAL MICROSOFT BLOG (Jan. 23, 2023), https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/.
[72] Reuters, *OpenAI may leave the EU if regulations bite - CEO*, REUTERS (May 24, 2023), https://www.reuters.com/technology/openai-may-leave-eu-if-regulations-bite-ceo-2023-05-24/.
[73] Supantha Mukherjee & Martin Coulter, *GhatGPT-maker OpenAI says has no plans to leave Europe*, REUTERS (May 26, 2023), https://www.reuters.com/technology/openai-has-no-plans-leave-europe-ceo-2023-05-26/.

of regulation to mitigate the most severe risks, Meta recently announced it will not release its upcoming multimodal Llama in the EU due to "the unpredictable nature of the European regulatory environment".[74]

Furthermore, one cannot dismiss the possibility that the leading AI players advocate for regulation to mask calculated self-interest. By speaking out in favor of regulation, those leaders can position themselves as the interlocutors of discussions between industry leaders and legislators and hope to influence laws, so that they serve, or at least do not thwart, their own interests.[75] Specifically, the stance of companies advocating for licensing could be seen as a strategy by established firms to secure market foreclosure to their advantage. By supporting regulatory measures, these firms may seek to create barriers that protect their dominant positions and hinder new competitors from entering the market.

1.3.2. Calls for international initiatives

In addition to government efforts, representatives from some AI companies have underscored the importance of global initiatives. For instance, Microsoft Vice Chair and President Brad Smith has promoted "multilateral public-private partnerships" to ensure AI governance at the international level.[76] According to Smith, international cooperation could "serve as an interim solution before regulations such as the AI Act come into effect," but also would help to build "a common set of shared principles that can guide both nation states and companies alike."[77] Moreover, "there is an opportunity for the European Union, the United States, the other members of the G7 as well as India and Indonesia, to move forward together on a set of shared values and principles." A "multilateral framework" is essential to harmonize various national laws, ensuring that "an AI system certified as safe in one jurisdiction is also recognized as safe in another." Smith gives the example of the common safety standards established by the International Civil Aviation Organization. Specifically, in Smith's view, an international code should be established, which would build on the principles for trustworthy AI developed by the OECD (*see section 6.2.1.*), provide a mechanism for AI developers to attest to the safety of their systems against internationally agreed-upon standards, and promote innovation and access by facilitating mutual recognition of compliance and safety across borders.

In October 2023, a cohort of leading industry voices, including co-founders of Anthropic, Inflection, DeepMind, and LinkedIn, as well as Eric Schmidt,[78] the former CEO of Google, jointly proposed an International Panel on AI Safety (IPAIS). The body, modeled on the Intergovernmental Panel on Climate

---

[74] Meta has also suspended plans to launch its AI assistant in the EU (for the EU framework see section 5.1) and has paused the deployment of its generative AI tools in Brazil (for the Brazilian framework see section 5.4.1), both in response to regulatory concerns. Jess Weatherbed, *Meta won't release its multimodal Llama AI model in the EU*, THE VERGE (July 18, 2024), https://www.theverge.com/2024/7/18/24201041/meta-multimodal-llama-ai-model-launch-eu-regulations.

[75] Deepa Seetharaman, *Efforts to Rein In AI Tap Lesson From Social Media: Don't Wait Until It's Too Late*, WALL ST. JOURNAL (July 17, 2023), https://www.wsj.com/articles/efforts-to-rein-in-ai-tap-lesson-from-social-media-dont-wait-until-its-too-late-d6d3fb49.

[76] Brad Smith, *Advancing AI governance in Europe and internationally*, MICROSOFT EU POLICY BLOG (June 29, 2023), https://blogs.microsoft.com/eupolicy/2023/06/29/advancing-ai-governance-europe-brad-smith/.

[77] *Id.*

[78] Mustafa Suleyman & Eric Schmidt, *Mustafa Suleyman and Eric Schmidt: We need an AI equivalent of the IPCC*, FINANCIAL TIMES (Oct. 18, 2023), https://www.ft.com/content/d84e91d0-ac74-4946-a21f-5f82eb4f1d2d.

Change,[79] would be "an independent, expert-led body empowered to objectively inform governments about the current state of AI capabilities and make evidence-based predictions about what is coming."[80]

Such a body, the proposal's advocates argued, is needed to improve lawmakers' "basic lack of understanding about what AI is" and curtail the impulse to overregulate it. The proponents also argued that IPAIS's independence, internationality, and narrow focus on "establishing a deep technical understanding of current capabilities and their improvement trajectories" would guarantee impartiality. And that impartiality would allow the IPAIS, in a global fashion, to effectively set standards and "shape protocols and norms" around transparency. The IPAIS was proposed not as an alternative to enforceable, legal mechanisms, but as a precursor to regulation and a way to provide neutral, high-quality information to decision-makers, to give them a realistic view of risks and avoid overregulation.[81]

Finally, industry representatives joined experts and researchers in calling for the drafting of an international AI treaty to mitigate AI risks and ensure that AI benefits all of humanity.[82] The proposed treaty should include global compute thresholds to regulate AI model training, a collaborative AI safety lab, the limitation of capabilities within safe limits, and an international compliance commission to monitor adherence to the treaty.

1.4. Purpose and structure

In the context described above, this report aims to evaluate existing and ongoing initiatives in the governance and regulation of generative AI, including approaches of self-regulation, co-regulation, and traditional government regulation. It begins by examining the primary risks associated with generative AI and the individual or collective measures AI companies have implemented to mitigate these risks. The report then reviews legislative frameworks being adopted or considered in different regions of the world.

While this report addresses the current state of technology, its risks, and industry practices, it is not intended to serve as a comprehensive state-of-the-art review. For a detailed overview of generative AI, the *International Scientific Report on the Safety of Advanced AI*, released prior to the Seoul summit in May 2024,[83] is recommended. That report provides a current, science-based understanding of the safety of advanced AI systems, particularly general-purpose AI systems.

The present report is structured as follows: Chapter 2 provides a general presentation of generative AI technology. Chapter 3 explores the main risks and challenges posed by generative AI. Chapter 4 details the

[79] The Intergovernmental Panel on Climate Change (IPCC) is the United Nations' body for assessing the science related to climate change. Composed exclusively of scientists from over 50 countries, it aims to provide neutral and reliable information to decision-makers. *See* IPCC, https://www.ipcc.ch/ (last visited Apr. 27, 2024).
[80] Suleyman & Schmidt, *supra* note 78.
[81] *Id.*
[82] URGING AN INTERNATIONAL AI TREATY: AN OPEN LETTER, https://aitreaty.org/ (last visited June 1, 2024).
[83] The current version remains an interim report, with the final version to be published before the AI Summit that will take place in France in 2025. *See* Bengio et al., *International Scientific Report supra* note 7.

individual and collective measures taken by AI companies to mitigate these risks. Chapter 5 offers an in-depth analysis of the main regulatory frameworks governing generative AI. Chapter 6 discusses current initiatives and ongoing efforts at the international level. Chapter 7 summarizes the key insights and overall findings.

Chapter 2 Generative AI: The technology and supply chain

Despite its success and widespread use, the term "generative AI" encompasses sophisticated technology and a complex, often opaque supply chain. Therefore, it is essential to clarify the nature of generative AI and its technical characteristics. This chapter will begin by defining generative AI (section 2.1), followed by a brief overview of the main stages in developing a generative AI model (section 2.2). Finally, it will highlight the key characteristics of the current supply chain (section 2.3).

2.1. What is generative AI?

Generative AI models are a category of deep-learning models that are "trained" on extensive datasets and that can then be directed to *generate content* based on the data on which they have been trained. Generative artificial intelligence (generative AI or GenAI) is capable of generating new content for users in a variety of formats, including text, images, sounds, videos, and more. That being said, it is essential to precisely define and understand the various terms associated with generative AI.

2.1.1. Artificial Intelligence

A deep exploration into generative AI naturally prompts the initial question of how it stands apart from artificial intelligence (AI) in general. In the context of this report, the definition of "AI" is one recently reformulated by the Organisation for Economic Co-operation and Development (OECD).[84] This updated definition describes an AI system as "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." It specifies that "different AI systems vary in their levels of autonomy and adaptiveness after deployment." These definitions are accompanied by explicit illustrations.[85]

The OECD also provides a definition of "model," which it distinguishes from "system." The OECD defines an "AI model" as "a core component of an AI system used to make inferences from inputs to produce

[84] Stuart Russell et al., *Updates to the OECD's definition of an AI system explained*, THE AI WONK (Nov. 29, 2023), https://oecd.ai/en/wonk/ai-system-definition-update.

[85] An "Explanatory Memorandum" published in March 2024 expands on each new word of the revised definition and notes that, despite the extensive work in defining an "AI system," there may, nevertheless, be additional criteria to "narrow or otherwise tailor the definition when used in a specific context." OECD, *Explanatory memorandum on the updated OECD definition of an AI system*, OECD ARTIFICIAL INTELLIGENCE PAPERS, NO. 8, OECD PUBLISHING, PARIS (2024) at 9, https://www.oecd-ilibrary.org/docserver/623da898-en.pdf.

outputs."[86] The concept of "inference" generally refers to the process by which a system generates an output from its inputs, typically occurring after deployment. The illustration below also shows that AI is largely data-driven. That is, AI "infers" or "learns" patterns that are used to generate its outputs from data. This process is commonly referred to as "training," and the datasets are referred to as "training data."

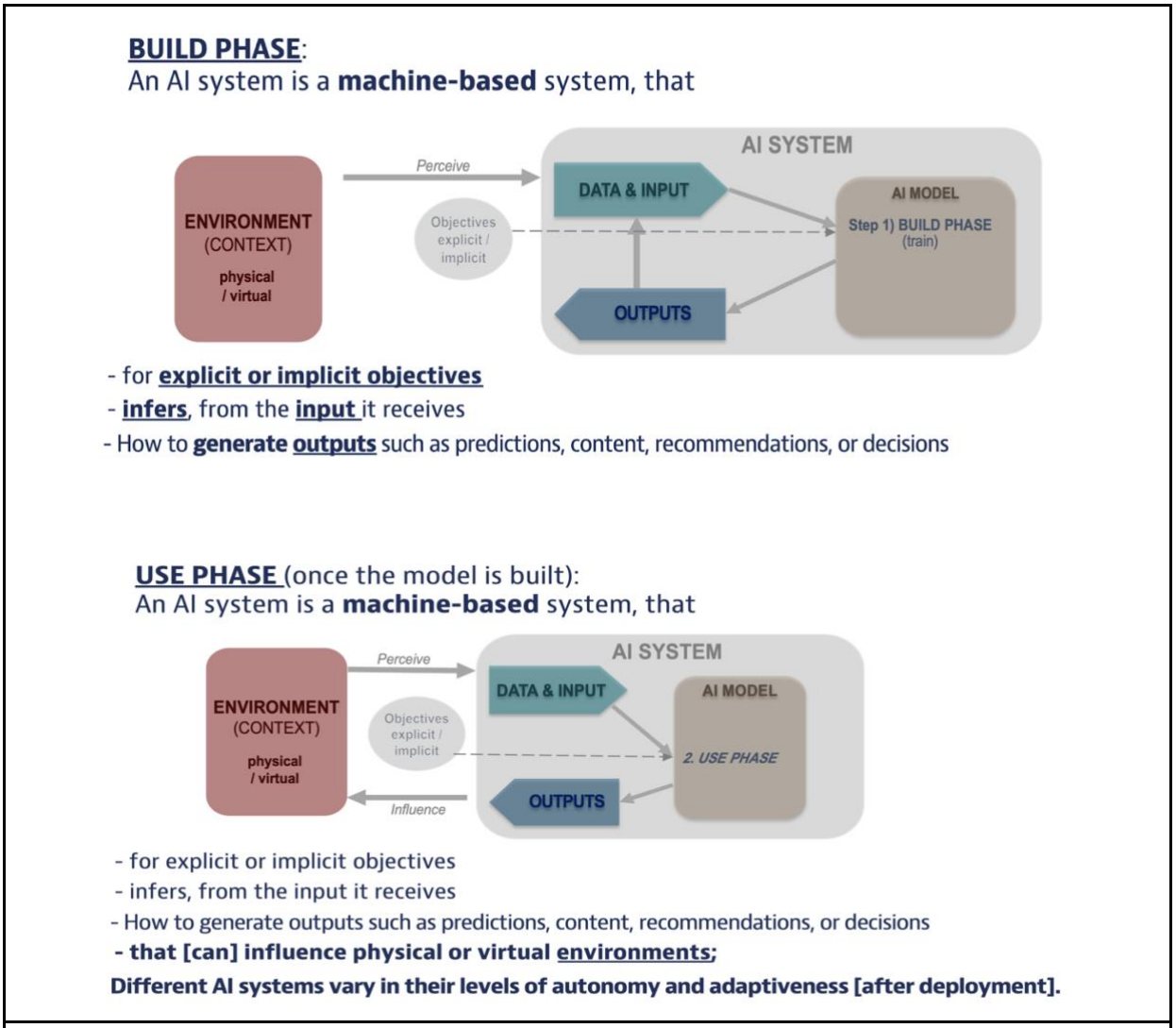


Figure 1. OECD's illustration of its definition of AI
Source: Stuart Russell et al., *Updates to the OECD's definition of an AI system explained*, THE AI WONK (Nov. 29, 2023), https://oecd.ai/en/wonk/ai-system-definition-update.

[86] Marko Grobelnik et al., *What is AI? Can you make a clear distinction between AI and non-AI systems?*, THE AI WONK (March 4, 2024), https://oecd.ai/en/wonk/definition.

In sum, an AI *model* is a program trained on a large set of data with the ability to identify patterns in that data in order to produce relevant outputs in response to inputs without the need for human intervention.[87] Although models represent a decisive part of development, they do not operate autonomously but are integrated into AI *systems*.[88] An AI system is typically built by combining one or more models. It also encompasses additional elements, such as a user interface, which are essential for its operation and interaction with humans.[89] Finally, for end use, AI *applications*, such as AI chatbots, leverage the capabilities of AI systems to perform specific tasks or solve a specific problem.

AI outputs can be of different kinds: predictions, recommendations, decisions, or content.[90] Generative AI refers to AI models and systems that are designed to fabricate new data based on the patterns, structures, and characteristics identified in the training data. In contrast, discriminative models are those that can discriminate between types (also known as "classes") of data, making them particularly useful for classification tasks. For example, discriminative models can determine whether an image contains a cat versus a dog, or whether an email is spam or not spam. Although discriminative AI systems may be used to fabricate new data, generative AI systems (i) are typically designed with this purpose in mind and (ii) generate outputs of the same modalities used to train the model. Within this framework, AI-generated data can take various forms, such as images, text, audio, videos, and computer code.[91]

2.1.2. Generative Artificial Intelligence

The field of generative AI is relatively new and continues to rapidly evolve, with a broad spectrum of terminology that lacks precise definitions or well-defined limits. Therefore, it becomes essential to elucidate the meanings of several key terms, before presenting the main use cases of generative AI.

2.1.2.A. Terminology

[87] Intuitively, this language evolved from the goal of AI being to "learn" a "model" of real-world concepts or processes. *See* IBM, *What is an AI model?*, https://www.ibm.com/topics/ai-model (last visited May 4, 2024).

[88] "AI models can be thought of as the raw, mathematical essence that is often the 'engine' of AI applications. An AI system is an ensemble of several components, including one or more AI models, that is designed to be particularly useful to humans in some way. For example, the ChatGPT app is an AI system. Its core engine, GPT-4, is an AI model," Bengio et al., *International Scientific Report* at 16.

[89] Recital 97 of the EU AI Act provides that "Although AI models are essential components of AI systems, they do not constitute AI systems on their own. AI models require the addition of further components, such as for example a user interface, to become AI systems. AI models are typically integrated into and form part of AI systems." Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). PE/24/2024/REV/1. OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj.

[90] *Background: What is a Generative Model?*, GOOGLE MACHINE LEARNING EDUCATION, https://developers.google.com/machine-learning/gan/generative (last visited Apr. 14, 2024).

[91] Exec. Order. 14,110 § 3(p), 88 Fed. Reg. 75191 (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, (defining generative AI models as "the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content").

Different expressions are commonly used, which are not necessarily synonymous and must be precisely defined.

1) Foundation models

Generative AI systems are primarily built upon "foundation models,"[92] a term that describes a class of AI models that provide foundational capabilities upon which other applications can be built. The term "foundation model" was coined in 2021 to fill a void in describing what scholars see as a "paradigm shift"[93] toward AI models "trained on broad data (generally using self-supervision at scale) that can be adapted or fine-tuned to a wide range of downstream tasks."[94]

Foundation models trace their roots to the early 2010s,[95] a period of AI development known as the "deep learning era" that began in roughly 2010 and continues to this day. The deep learning era was jump-started by significant progress in image classification,[96] a trend that has yielded substantial advances in AI.[97] The period also witnessed the emergence of a movement where large corporations, leveraging their extensive resources, released models of unprecedented scale. This began with the release of the breakthrough AlphaGo model in 2016 by Google DeepMind, which combined deep learning with reinforcement learning to master the game of Go. Advances in AI were further accelerated by the introduction of a deep-learning architecture known as the "transformer" in a 2017 paper.[98]

The transformer architecture marked a significant turning point for deep learning, particularly in the areas of natural language processing and computer vision. It enabled a huge leap in the amount of data that AI models could leverage and resulted in increased performance. This, in turn, enabled foundation models to suddenly possess the capacity to process vast and diverse volumes of unstructured data and undertake an extensive array of tasks. The two most popular types of transformers are *generative pre-trained*

---

[92] Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, arXiv (July 12, 2022), https://arxiv.org/pdf/2108.07258; *see also* Stanford University, *Ecosystem Graphs for Foundation Models*, https://crfm.stanford.edu/ecosystem-graphs/index.html?mode=table (last updated Mar. 27, 2024).

[93] *Id.*

[94] Rishi Bommasani et al., *Reflections on Foundation Models*, STAN. U. HUMAN-CENTERED ARTIFICIAL INTELLIGENCE (Oct. 18, 2021), https://hai.stanford.edu/news/reflections-foundation-models.

[95] Jaime Sevilla et al., *Compute Trends Across Three Eras of Machine Learning*, 2022 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN) (Mar. 9, 2022), https://arxiv.org/pdf/2202.05924.pdf.

[96] This work traces trends in the scale of compute resources used to train machine-learning models during the pre-deep learning (1952–2010), deep learning (2010–), and large-scale eras (2015–): Sevilla et al., *supra* note 95. The genesis of synthetic imagery can be traced back to 2014, rooted in the contributions of Ian Goodfellow et al., *Generative Adversarial Networks*, arXiv (June 10, 2014), https://arxiv.org/pdf/1406.2661, which set the stage for the capabilities exhibited by contemporary models. Due to swift advancements, synthetic visuals frequently challenge human discernment, appearing as genuine photographs.

[97] Sevilla et al., *supra* note 95.

[98] Ashish Vaswani et al., *Attention Is All You Need*, arXiv (Sept. 30, 2017), https://arxiv.org/pdf/1706.03762.

transformers (GPT) and *bidirectional encoder representations from transformers* (BERT).[99] OpenAI has used GPT to develop GPT-3 and GPT-4, while Google has refined BERT to develop Bard (now called Gemini).

Foundation models are typically described as possessing three main characteristics:

1. They require a **vast amount of data and computational resources** for their development. They are trained on a very large quantity of data, often collected on the internet by web scraping, and constructed on an enormous scale, comprising billions of adjustable parameters.
2. They possess the ability **to be adapted or fine-tuned**[100] **to suit a variety of specific downstream tasks**.[101] For instance, OpenAI's GPT-4 model can power chatbots that converse with users or assist in more specialized tasks, like performing content moderation on social media platforms.[102]
3. They exhibit a **high degree of complexity**, which makes it very difficult to understand how they operate. Specifically, they may acquire capabilities that extend beyond the developers' initial design objectives.[103]

Foundation models are not necessarily generative; they do not always produce or create new content. They can be used for non-generative tasks, like classification and information extraction. An example of this is CLIP (Contrastive Language–Image Pre-training), which excels at associating images and text (captions).[104] While it is foundational in the sense that it has a diverse array of downstream uses, it is not generative.

2) General-Purpose AI models and systems

The concept of "general purpose AI" models (GPAI) is typically used as a synonym for "foundation models." The term "general purpose" is indicative of the models' abilities to be adapted to a variety of tasks outside of those for which they were specifically trained.[105] This expression is sometimes preferred to the term "foundation models," or even "generative AI." For instance, the recent *International Scientific Report*

---

[99] Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv (May 24, 2019), https://arxiv.org/pdf/1810.04805.

[100] While it is possible to use foundation models without substantial modification, they are generally fine-tuned. Fine-tuning is the process of adding context-specific training.

[101] The term "foundation models" emphasizes their primary role: They are foundational and can be adapted to create many task-specific models. *See* Carlos Ignacio Gutierrez et al., *A Proposal for a Definition of General Purpose Artificial Intelligence Systems*, SSRN (Oct. 5, 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4238951.

[102] *See* Lilian Weng et al., *Using GPT-4 for content moderation*, OPENAI BLOG (Aug. 15, 2023), https://openai.com/blog/using-gpt-4-for-content-moderation.

[103] Bommasani et al., *supra* note 92.

[104] The CLIP algorithm "combines an image editor and a text editor to predict the correct pairings of a batch of image, and text training examples," Devlin et al., *supra* note 98.

[105] *See* Gutierrez et al., *supra* note 101.

on the Safety of Advanced AI specifically focuses on general-purpose AI models.[106] The report considers "an AI model to be general-purpose if it can perform, or can be adapted to perform, a wide variety of tasks." It also considers an AI system to be general-purpose "if it is based on a general-purpose model," but also "if it is based on a specialized model that was derived from a general-purpose model."[107] According to the report, "a model or system does not need to have multiple modalities, like speech, text, and image, to be considered general-purpose. Instead, AI that can perform a wide variety of tasks within specific domains, like structural biology, also counts as general-purpose."

3) Large models

Expressions like "Large Generative AI Models" (LGAIMs) [108] or "Large Language Models" (LLMs)[109] are also prevalent. The use of the expression "large models" emphasizes the fact that some AI models have a "large" number of parameters to make the models more robust and perform better for new, unseen data. Large models are related to foundation models in two ways: (i) because they have many parameters, they require significant data and resources to train, therefore aligning with the first characteristic of foundation models (requiring vast amounts of data); and (ii) because of their size, they align with the third characteristic of foundation models (high degree of complexity). When these large models are designed with the goal of being "general-purpose," they are referred to as foundation models.

Within this framework, the specific type of data employed in training a large model determines its functional "mode."[110] For instance, large models can be:

- **Language Models**: Language models, such as OpenAI's GPT-4o or Anthropic's Claude 3, are trained extensively on text data and generate output text that resembles human-generated text. These text-based models are usually referred to as Large Language Models (LLMs).
- **Text-to-Image Models**: Text-to-Image models, such as DALL·E 3, Stable Diffusion-3 or Midjourney, are trained on images and their corresponding textual descriptions to alter existing images or produce new images based on users' text prompt describing the desired image.
- **Audio Models**: Audio models, such as Google DeepMind's WaveNet, are trained on audio data and can be used for tasks like speech recognition, speech generation, or music generation.

---

[106] Bengio et al., *International Scientific Report supra* note 7 at 16.

[107] *Id.*

[108] Eduardo C. Garrido-Merchán et al., *ChatGPT is not all you need. A State of the Art Review of large Generative AI models*, arXiv (Jan. 11, 2023), https://arxiv.org/pdf/2301.04655; Deep Ganguli et al., *Predictability and Surprise in Large Generative Models*, PROCEEDINGS OF THE 2022 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (Oct. 3, 2022), https://arxiv.org/pdf/2202.07785; Philipp Hacker, et al., *Regulating ChatGPT and other Large Generative AI Models*, PROCEEDINGS OF THE 2023 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (May 12, 2023), https://arxiv.org/pdf/2302.02337.

[109] Wayne Xin Zhao, et al., *A Survey of Large Language Models*, arXiv (Nov. 24, 2023), https://arxiv.org/pdf/2303.18223v13.

[110] National AI Advisory Committee, *FAQs on Foundation Models and Generative AI*, NAIC (Aug. 28, 2023), https://www.ai.gov/wp-content/uploads/2023/09/FAQs-on-Foundation-Models-and-Generative-AI.pdf.

- **Video Models**: Video models, such as OpenAI's Sora, are trained on video data and can be used for tasks such as action recognition or video content generation.
- **Multimodal Models**: These models are trained on more than one data type. They process data from various sources, such as video, images, speech, sound, and text. As a result, they are useful for tasks that require understanding and processing multiple types of information. Since they can process information from different modalities, they can produce various types of outputs. For example, Google DeepMind's Gemini is a multimodal model that can be prompted with images, text, code, and video. Gemini possesses the capability to comprehend a wide range of input formats, integrate diverse information types, and produce a wide range of outputs.[111] Another example is OpenAI's GPT-4o, which accepts any combination of text, audio, image, and video as input, and generates any combination of text, audio, and image as outputs.[112]

4) Frontier models

Although there is no universally accepted or official definition for the term "frontier model," it is widely used among industry professionals and policymakers to refer to a subcategory of very advanced foundation models. A recent publication defines frontier models as "highly capable foundation models that could have dangerous capabilities sufficient to pose severe risks to public safety and global security."[113] The Frontier Model Forum, an industry body launched by the leading AI companies focused on ensuring safe and responsible development of frontier AI models (*see section 4.2.2.*), defines frontier models as "large-scale machine-learning models that exceed the capabilities currently present in the most advanced existing models, and can perform a wide variety of tasks."[114]

In other words, frontier models refer to a classification of powerful models that offer capabilities that are either novel or that surpass those of existing foundation models. Examples of these capabilities include designing new biochemical weapons, producing highly persuasive personalized disinformation, or being so autonomous that humans lose control of them. There is no consensus on a specific criterion for classifying a given model as a "frontier model," so the computational power required to train a model is sometimes used as an approximate indicator of the model's capabilities (*see below section 3.2.6.B*).[115]

---

[111] *Gemini*, GOOGLE DEEPMIND, https://deepmind.google/technologies/gemini/#build-with-gemini (last visited Apr. 15, 2024).

[112] OpenAI, *Hello GPT-4o* (May. 13, 2024), https://openai.com/index/hello-gpt-4o/.

[113] Markus Anderljung, et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, arXiv (Nov. 7, 2023), https://arxiv.org/pdf/2307.03718; *see also* Markus Anderljung, et al., *Frontier AI Regulation: Safeguards Amid Rapid Progress*, LAWFARE (Jan. 4, 2024), https://www.lawfaremedia.org/article/frontier-ai-regulation-safeguards-amid-rapid-progress.

[114] OpenAI, *Frontier Model Forum* (July 26, 2023), https://openai.com/index/frontier-model-forum.

[115] Neil C. Thompson, et al., *The Computational Limits of Deep Learning*, arXiv (July 27, 2022), https://arxiv.org/pdf/2007.05558.

5) Generative AI models and systems

As previously said, generative AI models fabricate new data based on the patterns, structures, and characteristics identified in their training data. Generative AI models are often foundation models.[116] For example, GPT-4o is, at the same time, a foundation model, a large multimodal model, and a generative AI model. However, some generative AI models are *not* foundation models. Generative Adversarial Networks (GANs), an alternative to the transformer architecture, have been widely used to power photo filters and other image generation applications since their introduction in 2014. But while they are able to generate highly realistic images, GANs generally lack the characteristic of applicability across a wide variety of tasks that characterizes foundation models.[117]

The majority of generative AI users do not engage directly with a generative AI *model.*[118] Rather, they interact through an interface with a generative AI *system* that incorporates the model. For example, ChatGPT, a generative AI system developed by OpenAI, is built on top of their latest and most advanced generative AI model, which is GPT-4o. Generative AI models are often one component among multiple embedded and interoperating components of an entire system. For instance, they can be embedded into software, such as office applications (Photoshop, PowerPoint, etc.) or be the "building" block of other AI systems.

Generative AI systems usually work by responding to a relatively simple input "prompt," or user-specified instructions, such as a text sentence. The user prompt can instruct the generative AI system to craft artificial content, which can include text, computer code, images, audio, or videos. For instance, a user can provide a written description of a painting to an image generation model, and the model will create visual content based on that description. Inputs and outputs can indeed be multimodal: different models can take in or produce various output formats, including text, audio, video, or a combination of several types.

This report focuses on the regulation of generative AI models and systems. While the main generative AI models are relatively easy to identify, generative AI systems are numerous and diverse. GPT-4 and GPT-4o are widely regarded as the leading foundation models, but recent developments have introduced competitive alternatives. In December 2023, Google released Gemini Ultra, which is sometimes presented as more powerful than GPT-4.[119] In January 2024, Adept AI introduced Fuyu-Heavy, which is recognized as the third most-capable multimodal AI model at the time of release, following Gemini Ultra and GPT-4V

---

[116] Sara Migliorini, *"More than Words": A Legal Approach to the Risks of Commercial Chatbots Powered by Generative Artificial Intelligence*, EURO. J. OF RISK REGULATION (Feb. 29, 2024) at 1–18, https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/more-than-words-a-legal-approach-to-the-risks-of-commercial-chatbots-powered-by-generative-artificial-intelligence/4EB4DD9997211B81283EF7B34299E254.

[117] Elliot Jones, *Explainer: What is a foundation model?*, ADA LOVELACE INSTITUTE (July 17, 2023), https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/; George Lawton, *GAN vs. transformer models: Comparing architectures and uses*, TECHTARGET (Apr. 12, 2023), https://www.techtarget.com/searchenterpriseai/tip/GAN-vs-transformer-models-Comparing-architectures-and-uses.

[118] Katherine Lee et al., *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, JOURNAL OF THE COPYRIGHT SOCIETY OF THE U.S.A. (Mar. 1, 2024), https://arxiv.org/pdf/2309.08133.

[119] Sundar Pichai & Demis Hassabis, *Introducing Gemini: our largest and most capable AI model*, THE KEYWORD (Dec. 6, 2023), https://blog.google/technology/ai/google-gemini-ai/.

(GPT-4 with vision).[120] Anthropic released Claude 3 in March 2024, asserting that it surpasses GPT-4 and Gemini Ultra across various benchmarks.[121] In April 2024, Meta introduced the Llama 3 model, which features advancements in reasoning and instruction-following capabilities, allegedly surpassing its predecessors in performance.[122] And it is anticipated that OpenAI will release an even more capable model, GPT-5, later this year.

The main generative AI models considered in this study are listed in the following table.

FIGURE 2. Main generative AI models

Company	Generative AI Model[123]
Adept AI (US)	Fuyu-Heavy
Aleph Alpha (Germany)	Luminous
Anthropic (US)	Claude 3
Baidu (China)	Ernie 4.0
Cohere (Canada)	Cohere Command
Google (US)	Gemini, PaLM 2, BERT
Hugging Face (US-France)	BLOOM
Meta (US)	Llama 3
Mistral AI (France)	Mixtral
OpenAI (US)	GPT-4, GPT-4o
Stability AI (US)	StableLM/ Stable Code 3B
Technology Innovation Institute (Emirates)	Falcon 180B
X AI (US)	Grok-1

2.1.2.B. Use cases

The possibilities and use cases of generative AI systems—specifically those systems capable of generating text, images, video, or computer code—are rapidly expanding, accompanied by increasingly reliable

[120] Adept, *Adept Fuyu-Heavy, A new multimodal model* ADEPT.AI, (Jan. 24, 2024), https://www.adept.ai/blog/adept-fuyu-heavy/.
[121] Anthropic, *Introducing the next generation of Claude*, ANTHROPIC, (March 3, 2024), https://www.anthropic.com/news/claude-3-family.
[122] Meta, *Introducing Meta Llama 3: The most capable openly available LLM to date*, META, (Apr. 18, 2014), https://ai.meta.com/blog/meta-llama-3/.
[123] These models are identified on the basis of HAI's most notable model releases of 2023. *See* Stanford AI Index Report 2024 *supra* note 3 at 78–80.

tools.[124] The use of generative AI by organizations and individuals has become widespread.[125] Sixty-five percent of businesses report that their organizations regularly employ generative AI in at least one business function, and most frequently in marketing and sales as well as in product and service development.[126] Among the numerous and varied use cases, only a very few are mentioned here.

1) Chatbots, search engines, and social media

Generative AI underlies general-purpose chatbots, such as OpenAI's ChatGPT, Google's Gemini , Anthropic's Claude, or Microsoft's Copilot. Specifically, Microsoft Copilot is designed to enhance productivity across various Microsoft 365 applications, and is integrated with tools like Word, Excel, PowerPoint, Outlook, and Teams. But generative AI models are also being used to power more specialized chatbots. For example, GitHub Copilot, which is primarily developed by GitHub (a subsidiary of Microsoft), is tailored for developers and focuses on coding assistance. Vik is designed by Wefight, Inc., to address the anxieties and queries of patients diagnosed with breast cancer. It has helped improve medication adherence rates.[127] And AcademiBot, an AI-powered educational dialog system, provides personalized assistance to students.[128]

Search engines are also augmenting their search capabilities by integrating conversational generative AI models. Both Microsoft Bing and Google Search now offer features that use generative AI to summarize search results, with sources cited.[129] Furthermore, in 2023, major social networking platforms integrated generative AI tools into their user experience. For example, Meta launched the Meta AI in September 2023, an AI assistant available in apps like WhatsApp, Messenger, and Instagram that provides search engine and image generation capabilities.[130] Finally, generative AI can also be used by social networks for content moderation tasks. OpenAI has promoted the use of GPT-4 for developing content policies and making

---

[124] Competition & Markets Authority, *AI Foundation Models: Initial Report* (Sept. 18, 2023), https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf.

[125] Alex Singla et al., *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*, MCKINSEY (May 30, 2024), https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/.

[126] *Id.* The use cases listed by the McKinsey report include personalized marketing, content support for marketing strategy, sales lead identification and prioritization, design development, scientific literature and research review, accelerated early simulation/ testing, IT helpdesk chatbot, data management, and IT helpdesk AI assistant.

[127] Benjamin Chaix et al., *When Chatbots Meet Patients: One-Year Prospective Study of Conversations Between Patients With Breast Cancer and a Chatbot*, NATIONAL LIBRARY OF MEDICINE (May 2, 2019), https://pubmed.ncbi.nlm.nih.gov/31045505/.

[128] Alexander Fox et al., *Revolutionizing Student Engagement and Enrollment through Personalized, AI-Driven Dialog Systems in Higher Education*, PROCEEDINGS OF THE 55TH TECHNICAL SYMPOSIUM ON COMPUTER SCIENCE EDUCATION V.2 (Mar. 15, 2024), https://dl.acm.org/doi/10.1145/3626253.3635414.

[129] Srinivasan Venkatachary et al., *A new way to search with generative AI: An overview of SGE,* GOOGLE (Jan. 2024), https://static.googleusercontent.com/media/www.google.com/en//search/howsearchworks/google-about-SGE.pdf; Yusuf Mehdi, *Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web*, MICROSOFT (Feb. 7, 2023), https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/.

[130] Meta, *Introducing New AI Experiences Across Our Family of Apps and Devices*, META NEWSROOM (Sept. 27, 2023), https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/.

content moderation decisions.[131] Allegedly, this approach allows for more consistent labeling, speeds up the feedback loop for refining policies, and reduces the need for human moderators.[132]

2) Content creation

Generative AI models produce creative writing, articles, and other textual content. AI-generated content may be used in various sectors, such as education[133] or research.[134] In addition, the visual domain (in particular, images and videos) has witnessed substantial advancements.[135] In February 2024, OpenAI released Sora, an AI system that creates realistic videos. In the music industry, AI-based melody generators allow artists to create music either from scratch or by building on existing musical phrases.[136]

Generative AI has shown particular value in helping developers create computer code. Some models are specifically trained on various programming languages to facilitate faster code production and potentially introduce coders to new syntactic or structural patterns.[137] For example, GitHub Copilot, powered by GPT-4, assists developers by suggesting complete lines or blocks of code as they type, effectively reducing the amount of manual coding.[138]

---

[131] OpenAI, *Using GPT-4 for content moderation* (Aug. 15, 2023), https://openai.com/index/using-gpt-4-for-content-moderation/.

[132] Paul M. Barrett & Justin Hendrix, *Is Generative AI the Answer for the Failures of Content Moderation?*, TECH POLICY PRESS (Apr. 3, 2024), https://www.techpolicy.press/is-generative-ai-the-answer-for-the-failures-of-content-moderation/.

[133] David Baidoo-Anu & Leticia Owusu Ansah, *Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning*, SSRN (Apr. 13, 2023), https://ssrn.com/abstract=4337484; Tammy Pettinato Oltz, *ChatGPT, Professor of Law*, SSRN (Feb. 6, 2023), https://ssrn.com/abstract=4347630.

[134] Will Douglas Heaven, *AI for protein folding*, MIT TECH. REV. (Feb. 23, 2022), https://www.technologyreview.com/2022/02/23/1044957/ai-protein-folding-deepmind/; Žiga Avsec et al., *Effective gene expression prediction from sequence by integrating long-range interaction*, NATURE (Oct. 4, 2021), https://www.nature.com/articles/s41592-021-01252-x; Sandra Brasil et al., *Artificial Intelligence (AI) in Rare Diseases: Is the Future Brighter?*, NAT'L LIB. OF MEDICINE (Nov. 27, 2019).

[135] Anne-Sofie Maerten & Derya Soydaner, *From paintbrush to pixel: A review of deep neural networks in AI-generated art*, arXiv (Feb. 14, 2023), https://arxiv.org/pdf/2302.10913.

[136] Li-Chia Yang et al., *MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation*, arXiv (July 18, 2017), https://arxiv.org/pdf/1703.10847; Emilia Gómez et al., *Deep Learning for Singing Processing: Achievements, Challenges and Impact on Singers and Listeners*, arXiv (July 9, 2018), https://arxiv.org/abs/1807.03046.

[137] Erik Nijkamp et al., *CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis*, arXiv (Feb. 27, 2023), https://arxiv.org/pdf/2203.13474; Prathamesh Ingle, *Top Artificial Intelligence (AI) Tools That Can Generate Code To Help Programmers (2024)*, MARKTECHPOST (Mar. 14, 2024), https://www.marktechpost.com/2024/03/14/top-artificial-intelligence-ai-tools-that-can-generate-code-to-help-programmers/.

[138] Thomas Dohmke, *GitHub Copilot is generally available to all developers*, GITHUB (June 21, 2022), https://github.blog/2022-06-21-github-copilot-is-generally-available-to-all-developers/.

3) Creation of synthetic datasets

Generative AI can create synthetic datasets that can then be used to train models.[139] From a small dataset, a generative AI model can create a larger database that respects the statistical properties of the original sample. This is particularly valuable in areas where data are scarce or sensitive. In sectors like fraud detection[140] or network security,[141] AI-generated data are used to assist models in recognizing anomalies or deviations.

2.2. Developing generative AI models

Developing generative AI models involves a large number of tasks, ranging from initial design and conceptualization activities to tasks related to the collection and preparation of data for the training, enhancement, and eventual deployment of the model itself. The interactions among development activities are complex, and there is no single development process that is shared across all AI producers.[142] Similarly, different actors can be involved in various development tasks. However, certain features of generative AI development can be identified. While certain activities will almost always precede others (for instance, model *pre-training* is conducted before model *fine-tuning*), the sequence of completed tasks may vary. The following paragraph provides a rough outline of the key tasks involved in developing a generative AI model.

2.2.1. Machine-Learning techniques

The remarkable successes of generative AI models can be credited to contemporary advancements in machine learning. Machine-learning techniques produce models that have learned the patterns and relationships expressed in the training data. That data can consist of segments of words or audio, parts of images or video, or a combination of these modalities.

In supervised machine learning, models learn from "training data"—the labeled data initially fed into a machine-learning model—in order to make predictions. Labeled data consist of input data paired with the expected output, which acts as the ground truth for the model to learn from. Originally, these models learned how to perform classification tasks from examples that had already been classified (labeled) by humans.

---

[139] Ryan Lingo, *Exploring the Potential of AI-Generated Synthetic Datasets: A Case Study on Telematics Data with ChatGPT*, arXiv (June 23, 2023), https://arxiv.org/pdf/2306.13700.

[140] Yinan Cheng et al., *Downstream Task-Oriented Generative Model Selections on Synthetic Data Training for Fraud Detection Models*, arXiv (Jan. 3, 2024), https://arxiv.org/pdf/2401.00974.

[141] Thomas Gaffney, *Synthetic data generation: Building trust by ensuring privacy and quality*, IBM BLOG (Nov. 29, 2023), https://www.ibm.com/blog/synthetic-data-generation-building-trust-by-ensuring-privacy-and-quality/.

[142] Katherine Lee et al., *supra* note 118. *See also* the presentation in Competition & Markets Authority, *AI Foundation Models: Initial Review* (May 4, 2023), https://assets.publishing.service.gov.uk/media/64528e622f62220013a6a491/AI_Foundation_Models_-_Initial_review_.pdf.

For example, classifiers learned how to categorize data into distinct classes (e.g., dogs vs. cats) from data that have been human labeled (i.e., photos that humans had labeled "cat" or "dog"). Generative models go further to produce new data (outputs) that retain important patterns and relationships of the original data without necessarily being identical to the original data.

The emergence of foundation models is the latest stage in the evolution of machine learning. To develop a foundation model, developers train a deep learning algorithm using extensive amounts of raw, unstructured, and unlabeled data. Thanks to transformers, it has become increasingly efficient to train models on such large volumes of data. Previously, developers gathered and labeled data to train a model for a specific task. Now, a single model can be trained on a vast dataset and subsequently adapted to various tasks by fine-tuning with a small amount of labeled, task-specific data (*see below section 2.2.3.*). By removing the necessity to define a task upfront, transformers enabled the pre-training of large models on extensive raw data, facilitating significant growth in model size.[143]

During the pre-training phase, the model learns by identifying and encoding the patterns and relationships in the data. The pre-trained model is then able to perform a variety of tasks by "transferring" its learned knowledge to new, related contexts. This is possible thanks to "transfer learning," which involves applying the "knowledge" acquired from one task to a related task.[144]

2.2.2. Model pre-training

The presentation of pre-training offered here is necessarily brief and schematic. Nonetheless, it can be stated that, overall, the model pre-training process typically involves data selection and curation followed by the learning phase.

2.2.2.A. Data collection and curation

The development of state-of-the-art generative AI models necessitates large volumes of data. Model developers can create and curate datasets, but often, training datasets are curated by other parties. The "Foundation Model Development Cheatsheet"[145]—a guide "prepared by foundation models developers for foundation models developers"—lists the available datasets according to the type of data they include (English text; multilingual text; specialized text, such as legal texts; image-text pairs; read speech, such as audiobooks, etc.).

---

[143] Kim Martineau, *What is generative AI?*, IBM RESEARCH BLOG (Apr. 20, 2023), https://research.ibm.com/blog/what-is-generative-AI; Cole Stryker & Mark Scapicchio, *What is generative AI?*, IBM RESEARCH BLOG (March 24, 2024), https://www.ibm.com/topics/generative-ai.

[144] Transfer learning is an old methodology in AI, with origins from the 1990s. *See, e.g.*, LEARNING TO LEARN (Sebastian Thrun & Lorien Pratt eds., 1998).

[145] Shayne Longpre et al., *The Foundation Model Development Cheatsheet,* GITHUB (Feb. 29, 2024), https://github.com/allenai/fm-cheatsheet/commits/main/app/resources/paper.pdf.

Due to the volume required, the data employed for pre-training often come from publicly accessible sources, though proprietary data can also be used in some instances. For example, Common Crawl, an open repository of web crawl data,[146] is commonly used to train generative AI models. The data in the Common Crawl database have been collected from webpages since 2008 using tools called "web crawlers," which covertly extract information from websites without leaving any trace of their activity.[147] These data are then "cleaned" by applying various filters to remove offensive words and other undesirable content. Other sources of data include, among others, the Project Gutenberg Corpus, a compilation of over 50,000 books in the public domain;[148] Wikipedia; and public open-source GitHub repositories.

Some of the existing datasets provide metadata detailing the origin of their data samples, but this information is often lacking. The practice of using web scraping to compile training datasets for generative AI complicates the tracking of data provenance: AI companies and data aggregators frequently deploy automated bots to search the web for new or updated webpages, which are subsequently scraped to gather training data.

AI developers are often reluctant to disclose the specific sources of their training data. However, there are occasional exceptions. For instance, Meta disclosed that the dataset employed for the pre-training of its first Llama model consists of several sources of data: Common Crawl (67%); the C4 dataset (15%); GitHub (4.5%); Wikipedia (4.5%); arXiv (2.5%); books (4.5%); and StackExchange, a public question and answer website (2%).[149]

2.2.2.B. Learning process

Following data collection and curation, the data are *tokenized*, or transformed into a format suitable for the training process. Tokens are small fragments of a text or image that serve as the basic units of data that a model processes. For text models, tokens might correspond to a word or a fragment of a word. As such, a training dataset can contain billions of tokens.

Throughout the pre-training phase, the model acquires an understanding of the probabilistic relationships between each token and all other tokens in the dataset. In virtually all leading generative models, an algorithm known as the *attention mechanism* enables the model to discern which tokens offer contextual information about the meanings of others. The attention mechanism identifies the relevance of inputs according to the specific context of the query and assigns them different "weights" in the algorithm's calculation process. The model thus generates outputs by predicting the most likely token to fit in a given context.

---

[146] *See* COMMON CRAWL, https://commoncrawl.org/ (last visited May 4, 2024).

[147] Migliorini, *supra* note 116.

[148] Martin Gerlach & Francesc Font-Clos, *A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics*, arXiv (Dec. 19, 2018), https://arxiv.org/pdf/1812.08092.

[149] Hugo Touvron et al., *LLaMA: Open and Efficient Foundation Language Models*, arXiv (Feb. 27, 2023), https://arxiv.org/pdf/2302.13971.

2.2.3. Fine-tuning

After unsupervised pre-training with raw data to acquire general-purpose representations, generative AI models often undergo additional training to better align with specific tasks and user preferences.[150] While the model has already developed foundational capabilities from a large dataset, it is further trained on a new smaller, task-specific dataset to adapt it for a particular task or domain, such as law or medicine. This "fine-tuning" process involves adjusting the model's weights and parameters to improve its performance on tasks in the specific subject area while retaining the knowledge it gained during the general-purpose pre-training phase.[151]

Sometimes, the original developer of an AI model will fine-tune it. In other cases, when a model's parameters are publicly released, other developers will independently fine-tune the model for particular applications. These third-party developers can accomplish fine-tuning by using either a local version of the model or an application programming interface (API). AI companies that release their models on an API enable downstream developers to make their desired adjustments. Fine-tuning enables the adjustment of the model for specific applications without the need to train it from scratch, thereby saving time and computational resources.

Fine-tuning is achieved via two main techniques: supervised fine-tuning and reinforcement learning. Supervised fine-tuning focuses on enhancing the model's performance on specific tasks. Reinforcement learning entails training the model to perform tasks more effectively and efficiently.

2.2.3.A. Supervised fine-tuning

Supervised fine-tuning (SFT) is the process of taking a pre-trained model, which has already developed foundational capabilities from a large dataset, and training it further using a smaller, task-specific dataset.[152] This SFT dataset is highly curated, often involving human annotation.[153] For example, SFT datasets can consist of labeled medical information to help a model perform better diagnostics.

SFT can also be used to improve a model's ability to generate content in a specific style or format. *Instruction Tuning*, a subset of SFT which is often applied to chat models, helps a model produce outcomes that fit the preferred style and objectives of a human engaging in chat conversation. This involves providing the model with examples of questions or prompts that the application is expected to handle, along with the corresponding correct answers or responses in the desired format. For example, for conversational models, developers often introduce high-quality examples of conversational responses written by humans.

---

[150] Chunting Zhou et al., *LIMA: Less Is More for Alignment*, arXiv (May 18, 2023), https://arxiv.org/pdf/2305.11206.

[151] Dave Bergmann, *What is fine-tuning?*, IBM, (March 15, 2024), https://www.ibm.com/topics/fine-tuning#:~:text=Fine%2Dtuning%20in%20machine%20learning,models%20used%20for%20generative%20AI.

[152] Bergmann, *supra* note 151.

[153] Rachel Lim et al., *Customizing GPT-3 for your application*, OPENAI BLOG (Dec. 14, 2021), https://openai.com/index/customizing-gpt-3.

2.2.3.B. Reinforcement learning with feedback

The reinforcement learning process customizes a model's responses and behaviors to be more aligned with a human user's expectations or preferences.[154] The AI model is trained to broadly reproduce desirable behaviors by "learning" from feedback on its actions. For example, the model can be trained not to use certain offensive or discriminatory language, such as racist terms, when feedback penalizes output containing such terms. This helps ensure the model generates responses that align with ethical and socially acceptable standards.[155]

One common approach is using "**Reinforcement Learning from Human Feedback**" (RLHF) where humans provide feedback in the form of "annotations."[156] These annotations are used to train a "reward model," a separate AI model that exclusively learns to predict what kinds of outputs are preferred based on the human-annotated examples. This reward model is then used to score the outputs of the primary model. The reward model essentially teaches the primary model how to produce outputs that earn the highest score, i.e., rewards. For example, to "fine-tune" its GPT-3 model, OpenAI first used feedback in the form of human-generated output examples,[157] then had humans rank the outputs of the model.[158] RLHF is also used by Anthropic, Meta, and Mistral, among others.[159]

An alternative approach is "**Reinforcement Learning with AI Feedback**" (RLAIF) (*see also section 4.1.1.C.*). RLAIF incorporates feedback from another AI system, instead of humans, to guide the learning process of the model.[160] Another AI model is trained on a set of principles. It then scores, or ranks, the primary model's outputs according to the predefined principles, identifying why the response is harmful, unethical, or illegal. The revision process may involve several iterations to refine the outputs. Using AI feedback in place of human feedback reduces the reliance on human annotators, making the process more scalable and cost-effective.[161]

A prominent example of using reinforcement learning is with applications that prioritize "natural" dialogue between users and the model, such as ChatGPT or Claude. Generally, the data used for pre-training contain

---

[154] Amazon Web Services, *What is Reinforcement Learning?*, https://aws.amazon.com/what-is/reinforcement-learning/ (last visited July 21, 2024).

[155] Yuntao Bai et al., *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*, arXiv (Apr. 12 2022), https://arxiv.org/pdf/2204.05862.

[156] Nathan Lambert et al., *Illustrating Reinforcement Learning from Human Feedback (RLHF)*, (Dec. 9, 2022), https://huggingface.co/blog/rlhf; Deval Shah, *RLHF (Reinforcement Learning From Human Feedback): Overview + Tutorial*, (June 29, 2023), https://www.v7labs.com/blog/rlhf-reinforcement-learning-from-human-feedback.

[157] Lambert et al. *supra* note 156.

[158] *Aligning language models to follow instructions*, OpenAI, (Jan. 27, 2022), https://openai.com/research/instruction-following.

[159] Stanford's *AI Index* found that the number of foundation models using RLHF rose from 0 in 2021 to 16 by 2023. Stanford AI Index Report 2024 *supra* note 3.

[160] Ryan O'Connor, *RLHF vs RLAIF for language model alignment*, Assembly AI, (Aug. 22, 2023), https://www.assemblyai.com/blog/rlhf-vs-rlaif-for-language-model-alignment/.

[161] *Id.*

a relatively small share of high-quality conversational text. In this situation, the resulting pre-trained model is said to be "under-weighted" for conversational behavior. As a result, it tends to extend the idea or topic of the initial prompt in a monologic style resembling the books, news articles, academic journals, and other monologic sources it was trained on. Without altering the training data, reinforcement learning can be used to encourage more conversational behavior by teaching a model to emphasize (increase the weighting of) conversational data. This emphasis is accomplished by having the model produce multiple outputs that feedback from a human (RLHF) or another AI model (RLAIF) ranks according to how well the outputs embody the desired conversational style. Because the model is designed to pursue the "reward" of having its outputs ranked highly, this "feedback" helps the model "learn" to produce outputs that better resemble that desired conversational style.

The same techniques of reinforcement learning can be used to *discourage* biased, false, or harmful outputs. However, this approach can lead to *overcorrection*, resulting in the model consistently applying these biases even when unnecessary. For instance, the early version of Google Gemini's image generation feature has been noted to generate historically inaccurate images. Gemini produced images depicting women or people of color at historic events or in positions once held only by white men at that time. So, an image might show a black woman as a pope. This can be seen as an overcorrection trying to ensure inoffensive and neutral outputs.[162]

2.2.4. Model architecture

In the early phases of creating a foundation model, developers focused on designing and implementing its architecture. Architecture encompasses decisions about the model's size (the number of parameters) and its topology (the structure of the network).

Determining the model's size mainly involves determining the number of parameters[163] or weights it will include. "Weights" are the numerical values that determine the strength of neural connections within a neural network and, thereby, help determine a model's output. During the training process, these weights are adjusted to optimize the model's performance, helping it produce more accurate and useful outputs. Furthermore, the relationship between the size of the model and its performance is mediated by the model's topology. "Topology" refers to the arrangement of neurons and layers within the neural network and how they are interconnected. It determines how input data are processed and how different features are extracted and combined to generate outputs.

A dramatic increase in the size (number of parameters) of generative AI models has been a key driver of recently improved capabilities. In 2018, GPT used only 117 million parameters and, in 2019, GPT-2 had

[162] Prabhakar Raghavan, *Gemini image generation got it wrong. We'll do better.*, THE KEYWORD (Feb. 23, 2024), https://blog.google/products/gemini/gemini-image-generation-issue/.

[163] Parameters are the internal variables that machine-learning models fine-tune throughout their training to enhance prediction accuracy. In deep learning, these parameters are predominantly the weights allocated to the links between neurons.

1.5 billion parameters.[164] But in 2020, OpenAI pioneered advancements in the field with the introduction of GPT-3, a cutting-edge language model at the time, boasting 175 billion parameters.[165] The size of GPT-4 (2023) has not been officially released, but it is believed to be significantly larger, with estimates suggesting up to 1 trillion parameters. A few examples of generative AI models and their sizes are listed in the following table:

FIGURE 3. Examples of generative AI models and their sizes

Model	Year	Number of Parameters
GPT-1	2018	117 million
GPT-2	2019	1.5 billion
GPT-3	2020	175 billion
GPT-4	2023	Estimated 1.76 trillion
BLOOM	2022	176 billion
Gemini Nano-1	2023	1.8 billion
Gemini Nano-2	2023	3.25 billion
Gemini Pro	2023	50 trillion
Gemini Ultra	2024	175 trillion

2.2.5. Resources required for development

The development and operation of generative AI models demand significant resources. Experts emphasize that "the three key technical inputs to producing AI capabilities are data, algorithms, and compute," collectively known as the "AI triad."[166] Additionally, teams of specialized researchers and engineers are essential. This section focuses on three major necessary resources: large-scale datasets, significant computational capacity, and substantial financial investment.

2.2.5.A. Data

[164] OpenAI, *Better language models and their implications* (Feb. 2019), https://openai.com/index/better-language-models/.

[165] Tom Brown et al., *Language Models are Few-Shot Learners*, Conference on Neural Information Processing Systems (NeurIPS) (Dec. 2020), https://arxiv.org/pdf/2005.14165.

[166] Girish Sastry et al. *Computing Power and the Governance of Artificial Intelligence*, arXiv (Feb. 13, 2024), https://arxiv.org/pdf/2402.08797.

Data are the most essential resource for an AI model. The capabilities of these models rely heavily on both the quantity and quality of the training data. A large quantity of data provides a broader range of examples from which the model can learn. High-quality data are accurate, complete, consistent, timely, and diverse so that models trained on them are reliable and free from bias. As raw data are always "dirty," they must be cleaned and processed to make them understandable by an AI model. Securing extensive databases of high-quality data often demands significant time and financial investment. This situation benefits large technology companies that possess substantial amounts of usable data due to their diverse activities.

For now, leading models usually rely on vast volumes of low-cost training data. In the future, it is likely that AI developers will increasingly utilize higher-quality data, which may be more costly. High-quality data may allow "small" models to achieve competitive results, even if the size of the dataset is limited.[167] Despite the higher costs associated with significantly better quality data, *quality* data permit a decrease in the quantity of data needed. However, enhancing the quality of datasets demands resources to assess data quality, given that the market for data lacks transparency and prices do not always accurately represent data quality.[168]

2.2.5.B. Computational resources

Generative AI models require substantial computational resources, whether during the pre-training, fine-tuning, reinforcement learning, or operating (or "inference") phase.[169]

1) Computation required for training

Given the substantial size of generative AI models and the extensive volume of data needed for training, the models must generally be trained and operated using specialized hardware that can handle the computational demands. Excluding salaries and research costs, these computational resources represent the most significant expenses in model development. The computational expense of developing a particular model hinges on the model's number of parameters and the size of its training dataset. Together, these factors dictate the number of operations (floating-point operations per second: FLOPS)[170] required for training the model. They also influence the extent to which the processor[171] architectures—such as Tensor Processing Units (TPUs), Graphical Processing Units (GPUs) or Central Processing Units (CPUs)—are

[167] Xinyang Geng et al., *Koala: A Dialogue Model for Academic Research*, THE BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH BLOG (Apr. 3, 2023), https://perma.cc/9HUC-K9KC.

[168] Haifei Yu & Mengxiao Zhang, *Data pricing strategy based on data quality*, 112 COMPUTERS & INDUSTRIAL ENGINEERING 1, 1–10 (Oct. 2017), https://www.sciencedirect.com/science/article/abs/pii/S0360835217303509.

[169] The "inference phase" corresponds to the model use phase, i.e., the iterative process during which the model generates tokens one at a time from the prompt. Tri Dao et al., *Flash-Decoding for long-context inference*, STAN. U. HUMAN-CENTERED AI, https://crfm.stanford.edu/2023/10/12/flashdecoding.html (last visited May 4, 2024).

[170] FLOPS is the abbreviation for "floating-point operations per second," a unit for measuring a computer's speed, based on how many of a particular type of mathematical operation it can perform in a second.

[171] "Processor" is the central component of a computer that performs computational tasks.

utilized. These are specialized chips[172] designed to accelerate the large-scale computations necessary for training AI models.

The cost of the chips themselves is high, and rising demand and tensions on semiconductor markets are driving up the cost. The H100 model from world leader Nvidia, used for AI software deployment, is estimated at between $30,000 and $40,000 per chip.[173] OpenAI used 10,000 Nvidia V100 GPUs to train GPT-3, and it took about 21 days for Meta to train its first Llama model using 2,048 Nvidia A100 GPUs.[174] Hugging Face's BLOOM, a 176 billion-parameter open-source model, was trained for 3.5 months on 1.6 terabytes of text using 384 GPUs.[175]

2) Computation required for operation

Once the model is operational, its regular operation and the data centers needed to host it generate significant costs. Relatively speaking, inference has less computational cost than training and can be performed on cheaper hardware. However, the compute demands for inference may be especially high when operating a large-scale service with a high load of queries.[176] It is estimated, for example, that operating ChatGPT could cost up to $700,000 per day, with these costs also linked to the operation of the processors.[177] Of course, while training is a substantial investment that must be made ahead of seeing any return, the inference costs usually scale with revenue.

3) Financial capital

The training of generative AI models can take months and cost millions of dollars.[178] According to estimates by the Stanford Institute for Human-Centered Artificial Intelligence (HAI), the training costs of state-of-the-art AI models have grown dramatically in recent years, with OpenAI's GPT-4 requiring an estimated $78 million worth of compute to train and Google's Gemini Ultra costing $191 million.[179] Anthropic CEO

---

[172] "Chip" refers to the physical integrated circuit that houses one or more processors. TPUs and GPUs are specialized chips designed to train AI models.

[173] Kif Leswing, *Nvidia's latest AI chip will cost more than $30,000, CEO says*, CNBC (Mar. 19, 2024), https://www.cnbc.com/2024/03/19/nvidias-blackwell-ai-chip-will-cost-more-than-30000-ceo-says.html.

[174] CNBC, *ChatGPT and generative AI are booming, but the costs can be extraordinary*, (Apr. 17, 2023), https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html.

[175] Stas Bekman, *The Technology Behind BLOOM Training*, HUGGING FACE BLOG (July 14, 2022), https://huggingface.co/blog/bloom-megatron-deepspeed.

[176] Competition & Markets Authority, *supra* note 124.

[177] Erin Woo & Amir Efrati, *OpenAI's Losses Doubled to $540 Million as It Developed ChatGPT*, THE INFORMATION (May 4, 2023), https://www.theinformation.com/articles/openais-losses-doubled-to-540-million-as-it-developed-chatgpt; Aaron Mok, *ChatGPT Could Cost over $700,000 per Day to Operate. Microsoft Is Reportedly Trying to Make It Cheaper.*, BUSINESS INSIDER (April 20, 2023), https://perma.cc/NY9H-2CCA.

[178] Or Sharir et al., *The Cost of Training NLP Models: A Concise Overview*, arXiv (Apr. 19, 2020), https://arxiv.org/pdf/2004.08900; Lennart Heim, *Estimating PaLM's training cost*, LENNART HEIM: BLOG (Apr. 5, 2022), https://blog.heim.xyz/palm-training-cost/; Peter J. Denning & Ted G. Lewis, *Exponential Laws of Computing Growth*, COMMUNICATIONS OF THE ACM (Jan. 1, 2017), https://cacm.acm.org/research/exponential-laws-of-computing-growth/.

[179] Stanford AI Index Report 2024 *supra* note 3.

Dario Amodei stated in an April 12, 2024, interview that the current generation of advanced models being trained will approach $1 billion and generations of models he expects to become available in 2025 and 2026 will cost $5 billion to $10 billion.[180]

Given their substantial costs, the most advanced models are usually developed by major tech companies. In contrast, smaller companies, constrained by their limited resources, often depend on existing foundation models. Developing a model from a pre-trained foundation model requires significantly less investment. For example, EleutherAI trained its large model using the GPT-3 dataset, which amounted to a cost of $400,000.[181] Within this context, some people fear that only a few wealthy companies can dominate the AI field in the future, which may result in a dependent relationship between AI providers and downstream deployers and users (*see section 3.4.1.*).

2.3 The supply chain

The supply chain for generative AI models and systems is highly intricate, involving a variety of providers with interdependent relationships. This supply chain can be broadly represented as a continuum ranging from "upstream" providers to "downstream" deployers and users.[182] Within this framework, a significant distinction exists between "open-source" and "closed-source" models.

2.3.1. Upstream providers vs. downstream deployers and users

Upstream activities involve model development and distribution, while downstream refers to "the markets in which foundation models are deployed."[183] Upstream developers of AI models often design their products to be both general purpose (i.e., suitable for as wide a range of applications as possible without modification) and modifiable for particular use cases (e.g., where a downstream company seeks to fine-tune a model with its own data to tailor the model to its specific needs).[184]

---

[180] *Transcript: Ezra Klein Interviews Dario Amodei*, N.Y. TIMES: THE EZRA KLEIN SHOW (Apr. 12, 2024), https://www.nytimes.com/2024/04/12/podcasts/transcript-ezra-klein-interviews-dario-amodei.html.

[181] Will Douglas Heaven, *The open-source AI boom is built on Big Tech's handouts. How long will it last?*, MIT TECH. REV. (May 12, 2023), https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta/.

[182] Sabrina Küspert et al., *The value chain of general-purpose AI*, ADA LOVELACE INSTITUTE (Feb. 10, 2023), https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/.

[183] Competition & Markets Authority, *supra* note 124.

[184] *See generally* Aspen Hopkins et al., *AI Supply Chains Aren't AI Value Chains*, SUBSTACK (Jan. 19, 2024); Sarah H. Cen et al., *Three proposals for regulating AI*, SUBSTACK (Aug. 7, 2023), https://aipolicy.substack.com/t/on-ai-deployment-series.

Currently, the leading developers of generative AI models include Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI. These companies build the foundation models that power generative AI systems. Some of these developers also deploy this technology through their own websites or apps, allowing end users to access generative AI systems, such as chatbots. This dual role means that a single entity can be both a developer and a downstream deployer.[185] For example, GPT-4 is a foundation model *developed* upstream by OpenAI, but ChatGPT is an application *deployed* downstream by OpenAI and powered by GPT-4. This example shows that the straightforward distinction between upstream and downstream is complicated by the fact that some stakeholders perform both upstream and downstream roles.

In addition to managing their own first-party apps that provide end user access to their generative AI tools, generative AI developers also generally offer third parties the ability to deploy their own models and systems via application programming interfaces (APIs).[186] These third parties can use APIs to access, integrate, and customize the models for their specific use cases. For instance, since March 2023, ChatGPT is not only available as a first party AI service on OpenAI's own website, but it is also available to third parties via the API. Developers may offer these APIs either for a fee or for free. For example, Meta's Llama 3 open-source AI model is available for free download and use. Ultimately, AI developers can bundle these APIs with other services to create platforms that enable third parties to develop their own applications.

Within this context, generative AI is accessible in various formats, such as web-based applications for end users, APIs, or direct download.

- **Web-based applications for end users:** Various applications are directly accessible to consumers, small businesses, and other end users. For instance, in March 2024, Andreessen Horowitz published a list of the most popular generative AI applications[187] used by consumers (see Appendix I). Examples include:
    - OpenAI's ChatGPT, renowned for its human-like responses, particularly since the introduction of GPT-4 and GPT-4o.
    - DALL·E, also developed by OpenAI, is a versatile multimodal application that connects words to visual components, enabling it to generate images based on user inputs.
    - Midjourney generates images based on natural language inputs and creates high-quality images from straightforward text prompts.
    - Google's Gemini, a chatbot initially launched under the name Bard in February 2023, originally leveraged Google's LaMDA (Language Model for Dialogue Applications) to provide a versatile and collaborative AI service integrated within Google Search. It is now powered by the Gemini model.

[185] Deployers may be defined as "entities or individuals that implement and manage AI technologies in user-facing applications or services." Center for American Progress, *Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone*, POLICY COMMONS (Feb. 1, 2024), https://policycommons.net/artifacts/11319438/generative-ai-should-be-developed-and-deployed-responsibly-at-every-level-for-everyone/12205150/.

[186] The Software Alliance (BSA), *AI Developers and Deployers: An Important Distinction* (Mar. 6, 2023), https://www.bsa.org/files/policy-filings/03162023aidevdep.pdf.

[187] Olivia Moore, *The Top 100 Gen AI Consumer Apps*, ANDREESSEN HOROWITZ (Mar. 13, 2024), https://a16z.com/100-gen-ai-apps/.

    - Microsoft's Copilot, an AI-powered assistant included in Microsoft products, assists in drafting and editing documents, helps analyze data, and may automate routine tasks, such as scheduling meetings or managing emails.
- **Application programming interfaces (APIs):** APIs allow downstream actors to access and incorporate a model's capabilities into their own applications and services. OpenAI, for instance, offers APIs for its GPT-3.5 and GPT-4 models. Integration through an API allows downstream developers to develop a specific application, such as a chatbot, powered by the foundation model. Note that this requires the business to share input data with the foundation model provider.
- **Direct download:** Open-source foundation models can be downloaded from the original upstream provider and independently deployed by downstream entities. Some developers offer their services at different levels. For instance, Stability AI provides its open-source Stable Diffusion model for direct download but also allows external users to access DreamStudio, a web interface incorporating the Stable Diffusion model.

A recent McKinsey survey revealed that approximately half of the generative AI applications utilized by companies implementing generative AI solutions rely on off-the-shelf, publicly available solutions, with little to no customization, rather than tools customized with proprietary data and systems.[188]

Partnerships add additional layers of complexity to the supply chain. For instance, OpenAI has partnered with Microsoft since 2019. Microsoft built a supercomputer on its Azure infrastructure to provide OpenAI with exclusive use of the hardware and computational resources needed to train its models.[189] In September 2023, Anthropic announced a similar partnership with Amazon that makes Amazon Web Services (AWS) its primary cloud provider and provides Anthropic with access to Amazon's compute infrastructure.[190] In addition to the provision of upstream infrastructure, the close partnerships between leading generative AI developers and the established tech companies have a significant impact on downstream distribution. Anthropic's Claude is available to Amazon developers and enterprise customers via Amazon Bedrock. Microsoft has similarly integrated OpenAI and Mistral models into its Azure ecosystem.

FIGURE 4. Supply Chain actors: schematic overview[191]

	Actors	Services	Select Examples
AI Infrastructure	Company manufacturing hardware	Makes GPUs, TPUs, or CPUs	Nvidia (GPUs) Google (TPUs)

[188] Singla et al., *supra* note 125.

[189] Rob Waters, *Microsoft built a Supercomputer to power OpenAI's ChatGPT*, CYBERSECURITY CAREERS BLOG (Mar. 13, 2023), https://www.cybercareers.blog/2023/03/microsoft-built-a-supercomputer-to-power-openais-chatgpt/; Dina Bass, *Microsoft Strung Together Tens of Thousands of Chips in a Pricey Supercomputer for OpenAI*, BLOOMBERG (Mar. 13, 2023), https://www.bloomberg.com/news/articles/2023-03-13/microsoft-built-an-expensive-supercomputer-to-power-openai-s-chatgpt?sref=CY7BpvSE&embedded-checkout=true.

[190] Anthropic, *Expanding access to safer AI with Amazon* (Sept. 25, 2023), https://www.anthropic.com/news/anthropic-amazon.

[191] *See* Jones, *supra* note 117.

AI Infrastructure	Company building large-scale computing systems	Builds hardware systems to power for training and inference	Amazon Google Microsoft
AI Infrastructure	Company selling access to compute and storage	Provides cloud compute and storage	Amazon (AWS) Google (Google Cloud) Microsoft (Azure)
AI Infrastructure	Company collecting and processing data	Gathers data at scale and make them accessible to the public	Common Crawl (Common Crawl Corpus) Eleuther (ThePile) LAION (LAION-5B)
AI Modeling	Company developing AI models	Designs and trains AI models and makes them available for download free of charge	Eleuther (GPT-NeoX) Meta (Llama 3) Mistral (Mixtral) Stability.AI (Stable Diffusion)
AI Modeling	Company developing AI models	Designs and trains AI models and makes them available for a fee via API	Anthropic (Claude 3) ElevenLabs (Multilingual v2) Google (Gemini) OpenAI (GPT-4)
AI Modeling	Company developing AI models	Designs and trains AI models for their own proprietary use	Character (Character.ai) OpenAI (Voice Cloning)
Model Hosting	Company providing model repository	Hosts and makes models accessible to downstream users	Civitai Hugging Face
Model Hosting	Company providing combined model repository and cloud compute	Hosts and makes models available to downstream users via download or API access	Amazon (AWS) Hugging Face Google (Google Cloud) Microsoft (Azure) Perplexity (pplx-api)
Downstream Deployer	Company developing applications using AI models	Releases applications powered by models (which they may or may not have fine-tuned)	Character (Character.ai) GitHub (Copilot) Microsoft (Bing/Sidney) Notion (Notion.so) OpenAI (ChatGPT) SalesForce (Einstein)
Downstream User	End users using AI-powered applications	Use applications to generate text, images, videos, computer code	Individual users and businesses

2.3.2. Open-Source vs. Closed-Source release

Generative AI release strategies typically range from closed models, such as proprietary, commercial, or internal-use-only, to open-source models. Noncommercial open-source models provide users with access to both the weights and training methodologies.

The term “open source” as used to describe AI model and systems is borrowed from the expression “open-source software.” Open-source software is defined as “software designed to be publicly accessible—

meaning anyone can view, use, modify, and distribute the source-code—and that is released under an open-source license."[192] The features for an open-source license are very specific. They include free source code access, permission for modifications and derived works, and no discrimination against which fields or groups may use the software.[193]

In the field of AI, the release of an open-source AI model usually allows users to download, modify, and share the entire model or specific parts of it. Conversely, closed-source models, which are usually created within private companies, have limited accessibility. Their developers can use these closed-source models internally for their products or processes, or they can provide them to external parties for use under specific conditions.[194]

1) Open vs. Closed-Source is a continuum, not a binary

Despite the common perception that "open" is the opposite of "closed," the descriptor of "open source" as applied to AI systems has a variety of meanings, though they all indicate public accessibility. Researchers have argued against the characterization of open- versus closed-source AI systems as a dichotomy.[195] Instead, they advance the view that different attributes of release and access fall along a gradient of open to closed. Irene Solaiman has outlined six different approaches for an AI company to use when releasing AI models,[196] ranging from "fully closed" to "fully open."

- **Fully closed systems**: In "fully closed" models, all aspects and components of the system are inaccessible outside the developer organization or even within specific subsections of that organization. Examples include Google DeepMind's Gopher.[197]
- **Gradual/Staged release**: This variation involves releasing a system gradually over a set period. It allows developers time to monitor for malicious activities or to conduct research on potential harms. In 2019, OpenAI released its language model GPT-2 in four sizes, increasing the parameter count over nine months.[198]
- **Hosted access**: Users can access the model on the provider's servers with "surface-level" interaction that provides the users with only very limited ability to shape model behavior. OpenAI's ChatGPT is an example of hosted access.

---

[192] Elizabeth Seger et al., *Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*, arXiv (Sep. 29, 2023), https://arxiv.org/pdf/2311.09227.

[193] *See* Open Source Initiative, *The Open Source Definition* (Feb. 16, 2024), https://opensource.org/osd.

[194] Competition & Markets Authority, *supra* note 124.

[195] Rishi Bommasani et al., *Considerations for Governing Open Foundation Models*, STAN. U. HUMAN-CENTERED AI (Dec. 13, 2023), https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models.

[196] Irene Solaiman, *The Gradient of Generative AI Release: Methods and Considerations*, PROCEEDINGS OF THE 2023 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (Feb. 5, 2023), https://arxiv.org/pdf/2302.04844.

[197] DeepMind's Gopher is a large language model designed for natural language understanding and generation. It remains a closed model used primarily for research and internal projects. Jack Rae et al., *Language modelling at scale: Gopher, ethical considerations, and retrieval*, GOOGLE DEEPMIND (Dec. 8, 2021), https://deepmind.google/discover/blog/language-modelling-at-scale-gopher-ethical-considerations-and-retrieval/.

[198] Solaiman, *supra* note 196.

- **Cloud-based access or API access**: Some models are hosted on cloud platforms and accessible via an API. This level of access allows third parties to access the model and includes the flexibility to adjust some of its parameters to shape outputs without fundamentally altering its core structure. Downstream developers can use and adapt the model for specific applications, while the model is hosted and operated on the provider's servers. The extent of access allowed by each provider through an API varies significantly. It can range from highly restrictive, where only the output is shared, to more customizable arrangements that allow users to fine-tune the model for a specific task (cloud-based fine-tuning access). Via this method, the provider maintains control over the model and can monitor usage to detect and prevent abuse by users and developers. Notable examples of models distributed via API include OpenAI's GPT-4 and Anthropic's Claude 3, both of which are also available via hosted access.
- **Downloadable models**: Third parties may be allowed to download a model and deploy it on their systems, eliminating the need to share data with the model provider. Downloadable models may withhold certain key components, such as the training dataset. Some high-profile models that purport to be open source are arguably better categorized as downloadable. Meta's Llama 3 or Mistral AI's Mixtral are downloadable as fully trained models, but the model's training data and the code used to train it are not accessible.[199] Additionally, this approach does not necessarily grant full access to all users, as the model's size may limit the ability of some users to run it.
- **Fully open models**: In this iteration, all aspects of the models are accessible and downloadable. Most open models provide access to weights, code, and data without usage restrictions. However, the level of documentation detail and granularity may vary. Notable examples of open-source models include Google's BERT,[200] GPT-J,[201] RedPajama,[202] and StarCoder.[203] Hugging Face maintains an Open LLM Leaderboard that features numerous open-access models, further encouraging development.[204]

---

[199] Michael Nolan, *Llama and ChatGPT Are Not Open-Source: Few ostensibly open-source LLMs live up to the openness claim*, IEEE SPECTRUM (July 27, 2023), https://spectrum.ieee.org/open-source-llm-not-open.

[200] Jacob Devlin & Ming-Wei Chang, *Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing*, GOOGLE RESEARCH BLOG (Nov. 2, 2018), https://research.google/blog/open-sourcing-bert-state-of-the-art-pre-training-for-natural-language-processing/.

[201] GPT-J is an open-source language model developed by EleutherAI. It is available for anyone to download and use. *See* GPT-J, ELEUTHER AI https://www.eleuther.ai/artifacts/gpt-j (last visited June 15, 2024).

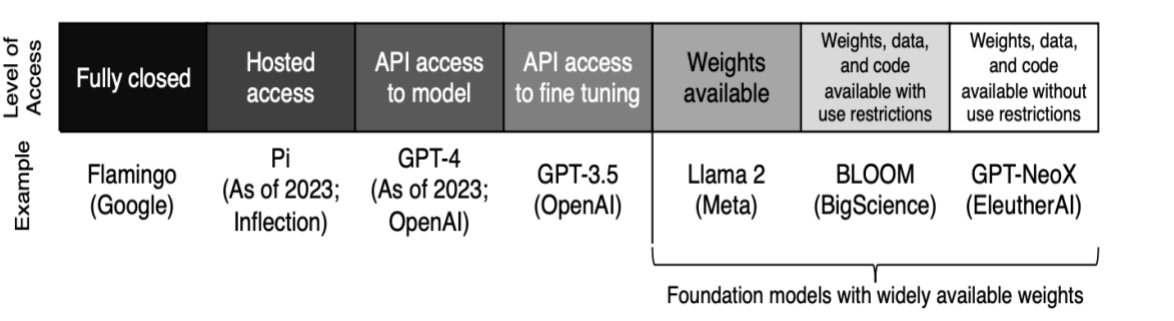
[202] The RedPajama project aims to create leading open-source large language models by reproducing and extending the LLaMA dataset, which contains over 1.2 trillion tokens. It is a collaborative effort involving several organizations, including Together, Ontocord.ai, ETH DS3Lab, Stanford CRFM, and Hazy Research. The dataset and models are fully open source, and can be accessed and downloaded through platforms like Hugging Face and GitHub. *See RedPajama, a project to create leading open-source models, starts by reproducing LLaMA training dataset of over 1.2 trillion tokens*, TOGETHER.AI, (April, 17, 2024) https://www.together.ai/blog/redpajama.

[203] Leandro von Werra and Loubna Ben Allal, *StarCoder: A State-of-the-Art LLM for Code*, HUGGING FACE (May 4, 2023) https://huggingface.co/blog/starcoder.

[204] *See* HUGGING FACE, https://huggingface.co/ (last visited June 20, 2024); BIG SCIENCE, https://bigscience.huggingface.co/ (last visited June 20, 2024); BIG CODE https://www.bigcode-project.org/ (last visited June 20, 2024).

In their modified version of Solaiman's figure (*see below figure 5*), Bommasani, et al.[205] describe the gradient as ranging from "fully closed" on one end of the spectrum to "models with widely available weights" on the other end, with "hosted access," "API access to model," and "API access to fine tuning" as intermediate levels of openness. Depending on the choices made by a provider's model, the release of open-source models may include the model's source code, its structural blueprint, and the data used for training, enabling others to duplicate the training procedure. It may also include the model's weights—its "knowledge"—allowing others to employ or adjust the model without undergoing their own initial training.

FIGURE 5. A revised spectrum of closed- to open-source models



Source: Rishi Bommasani et al., *Considerations for Governing Open Models*, STAN. U. HUMAN-CENTERED ARTIFICIAL INTELLIGENCE (Dec. 2023), https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

There is an ongoing debate about the criteria for true open-source AI, including issues of licensing, data transparency, and the accessibility of the models for research and commercial use.[206] For example, Meta's Llama 2 provided users access to the model's weights, evaluation code, and documentation.[207] However, Meta did not share the model's training data or the code used to train it, which made its classification as an open-source model questionable.[208] As for the recently released Llama 3 model, the company stated that its "final decision on when, whether, and how to open source will be taken following safety evaluations we will be running in the coming months."[209]

In any case, open-source models' usage may be limited by licensing terms that restrict their usage and distribution. Meta's initial Llama model was designated solely for research use, explicitly excluding commercial purposes. Recent trends suggest a change in direction. Meta's Llama 2, despite some

[205] Bommasani et al., *Considerations for Governing Open Foundation Models*, *supra* note 195.

[206] Ed Gent, *The tech industry can't agree on what open-source AI means. That's a problem.*, MIT TECHNOLOGY REVIEW (Mar. 25, 2024), https://www.technologyreview.com/2024/03/25/1090111/tech-industry-open-source-ai-definition-problem/.

[207] Meta Llama, https://ai.meta.com/llama/.

[208] Nolan, *supra* note 199.

[209] Meta, *Our responsible approach to Meta AI and Meta Llama 3* (Apr. 18, 2024), https://ai.meta.com/blog/meta-llama-3-meta-ai-responsibility/.

limitations, offered commercial licensing options.[210] Llama 3 is available for use, reproduction, distribution, and modification under a nonexclusive, worldwide, and royalty-free license.[211] However, Meta restricts the use of Llama 3's outputs for improving other large language models that are not derivatives of Llama 3. Additionally, commercial entities with more than 700 million monthly active users must seek a special license from Meta. These licensing terms raise important questions about the future control of open-source models, especially concerning potential constraints on their development and usage. In practice, it is particularly challenging to ensure adherence to these usage restrictions.

For its part, OpenAI has indicated that it is exploring the release of open-source models for commercial purposes but has not committed to a definitive timeline or specifics. The company balances the benefits of open access with the need to mitigate potential risks associated with AI misuse.[212]

2) Merits of closed-source and open-source models

The relative merits of open-source and closed-source models are a topic of ongoing debate (*see also section 3.2.6.A*) .[213] Some argue that "open sourcing allows more people to contribute to AI development processes and enables large-scale collaborative efforts. The idea is that more expertise, more diverse perspectives, and simply more human creativity and hours put into AI development will drive innovation in new and useful downstream integrations, advance AI safety research, and help push forward the boundaries of AI capability."[214]

Proponents of open-source models tout their increasing ability to match the performance of established closed-source models. In 2023, a leaked memo,[215] purporting to be from a "researcher within Google" and shared by an anonymous individual on a public Discord server, stated that open source makes it possible to efficiently fine-tune large AI models for specific tasks at a reduced cost. Since numerous developers are now utilizing open-source models, thereby circumventing the significant expenses associated with developing a new model, big players, such as Google or OpenAI, are not "positioned to win" the AI race, the memo said. While the proprietary models developed by leading AI companies still hold a slight edge in terms of quality, "the gap is closing astonishingly quickly," the memo said , as "open-source models are faster, more customizable, more private, and pound-for-pound more capable."[216]

---

[210] While Meta's license makes Llama 2 free for many, it requires a license fee for any developers with more than 700 million daily users and disallows other models from training on Llama.

[211] Joseph Spisak, *Meta Llama 3 Community License Agreement* (Apr. 18, 2024), https://github.com/meta-llama/llama3/blob/main/LICENSE.

[212] OpenAI, *OpenAI's comment to the NTIA on open model weights* (Mar. 27, 2024), https://openai.com/global-affairs/openai-s-comment-to-the-ntia-on-open-model-weights/.

[213] *See* Kevin Klyman, *How to Promote Responsible Open Foundation Models*, STAN. U. HUMAN-CENTERED AI (Oct. 3, 2023), https://hai.stanford.edu/news/how-promote-responsible-open-foundation-models.

[214] Seger et al., *supra* note 192.

[215] Dylan Patel & Afzal Ahmad, *Google "We Have No Moat, And Neither Does OpenAI,"* SEMIANALYSIS (May 4, 2023), https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.

[216] *Id.*

It is difficult to determine if applications fine-tuned from open-source models can truly compete with proprietary models and potentially unseat leading industry players. According to the 2024 AI Index from Stanford University, closed models outperform open ones, with a median performance advantage of 24.2%.[217]

2.3.3. Profitability models

As previously discussed (*section 2.3.2.1*), the methods for introducing generative AI models and systems to the market are highly diverse and range along a spectrum from fully closed to open-source models. Each of these market approaches aligns with distinct profitability models. Although the distinction between proprietary models and open-source models is not strictly binary, it is still possible to differentiate them in terms of monetization.[218]

1) Monetization of closed-source models

Developers of generative AI models may opt to use their models exclusively for their own benefit, integrating them into existing products and services to improve performance or introduce new features, rather than making them available to third parties. However, in most cases, generative AI developers offer access to their AI systems to end users through subscription models. Another monetizing option is a "freemium" approach, where access to the model is provided free of charge, but additional services or features are charged. For example, ChatGPT is a free internet service, similar to a search engine. As this report is written, this free version of ChatGPT is powered by GPT-3.5. However, OpenAI also offers a subscription plan for ChatGPT Plus, which is priced at $20 monthly. This plan includes improved features, such as access to GPT-4 and GPT-4o, higher usage limits, priority access during the peak times, and faster response times.

Generative AI model developers may also offer AI as a service for third parties that may use the model for their own needs or even incorporate the model into their products or services. To that end, third parties can access the model through a web interface (hosted access) or an API (cloud-based access). These services are usually billed based on usage (e.g., price per 1,000 or 1 million tokens of prompt data processed).[219] On the OpenAI API, downstream developers initially receive a set amount of complimentary credit to access or fine-tune the language model. Once this credit is exhausted, they have the option to purchase additional resources. Downloadable models may also be offered on a usage-based billing model.

2) Monetization of open-source models

When developers release their models as open source, they usually do not charge for access. While they have to invest substantial resources to develop their models, they do not directly earn profits from these

[217] Stanford AI Index Report 2024 *supra* note 3 chapter 2.8.

[218] Competition & Markets Authority, *supra* note 124.

[219] Louis Muswell, *OpenAI API Pricing: How Much Does It Cost?*, OPENAI (Apr. 7, 2024), https://openaidiscovery.com/openai-api-pricing.

products. However, they may adopt this strategy to encourage other developers and users to engage with and adopt their models. This can, in turn, lead to improvements in the quality of the models they use in other products or operations. Moreover, by promoting values of openness and collaboration, they can attract prospective employees in the highly competitive generative AI industry, where securing top talent is crucial.

Additionally, releasing models as open source can foster the growth of activities that complement the developers' AI products. For example, they may offer computational resources and support services to users of the open-source models. A possible option is to market software that facilitates interaction with or fine-tuning of the open-source model.

Certainly, by dropping the price that can be charged, open sourcing can undercut competitors that rely on charging for their models to sustain their operations. This problem is particularly acute when AI models are made available free of charge by large companies such as Meta, who can afford to do so because they have other profitable activities.

KEY TAKEAWAYS

- The term "generative AI" refers to a category of deep learning models that are "trained" on extensive datasets to identify patterns in the data. These models can subsequently be directed to generate content based on the patterns, structures, and characteristics they identified in their training data. Such content can be produced in various formats, including text, images, sounds, videos, computer code, and more. Generative AI *models* are integrated into generative AI *systems*, which include additional elements, like a user interface, that are essential for their operation and human interaction. Generative AI *systems* typically operate by responding to straightforward inputs, known as "prompts," which are user-specified instructions, such as a text sentence.

- The field of generative AI encompasses numerous technical terms that must be clearly defined and distinguished. Generative AI systems are primarily built upon "foundation models," a term that describes a class of AI models that provide foundational capabilities upon which other applications can be built. Foundation models, also known as "general purpose AI models" (GPAI), are trained on extensive datasets, can be adapted or fine-tuned for a range of specific downstream tasks, and exhibit a high degree of complexity. These models act as the foundational layers upon which more specialized models can be fine-tuned for specific tasks. Large AI models also require vast amounts of data and display a significant degree of complexity. They are foundation models when they are designed with the goal of being "general-purpose." Finally, the term "frontier model" refers to a subcategory of powerful models that offer capabilities that are either novel or that surpass those of existing foundation models.

- Developing generative AI models is a complex process that involves successive stages, ranging from data collection to fine-tuning. In very basic terms, the model pre-training process generally involves selecting and curating data, followed by a learning phase. Due to the large volume needed,

the data used for pre-training usually come from publicly accessible sources and are often web-scraped from the Internet. Following data collection and curation, the data are transformed into a format suitable for the training process. During the pre-training phase, the model learns by identifying patterns and acquires general-purpose representations. After pre-training, AI models are generally "fine-tuned." This involves further training on smaller, task-specific datasets to enable them to perform particular tasks. This is typically accomplished through supervised fine-tuning and reinforcement learning. Supervised fine-tuning involves training the model on a specific dataset to improve its performance on a particular task. Reinforcement learning customizes the model's responses and behaviors to better align with a human user's expectations and preferences.

- Developing an AI model requires highly advanced expertise and substantial resources, including data, computational power, and significant financial investment. Training generative AI models can take months and cost millions of dollars. The expenses associated with state-of-the-art AI models have surged dramatically in recent years. As a result, the most advanced models are typically developed by major tech companies. In contrast, smaller companies, constrained by limited resources, often rely on existing foundation models. This situation has led to concerns that a few wealthy companies may come to dominate the AI field, potentially creating a dependent relationship between AI providers and downstream deployers and users.

- The supply chain for generative AI models and systems is highly intricate, involving a variety of providers with interdependent relationships. This supply chain can be broadly represented as a continuum ranging from "upstream" providers to "downstream" deployers and users. A single entity can function as both an upstream developer and a downstream deployer. Generative AI developers often manage their own first-party applications, providing end users access to their tools, while also offering third parties the ability to deploy their own models and systems via application programming interfaces (APIs). Some developers make their generative AI models available for direct download. Tech giants play a pivotal role in this chain, frequently assuming multiple roles, such as developers of foundation models, providers of cloud solutions, and suppliers of AI applications to end users.

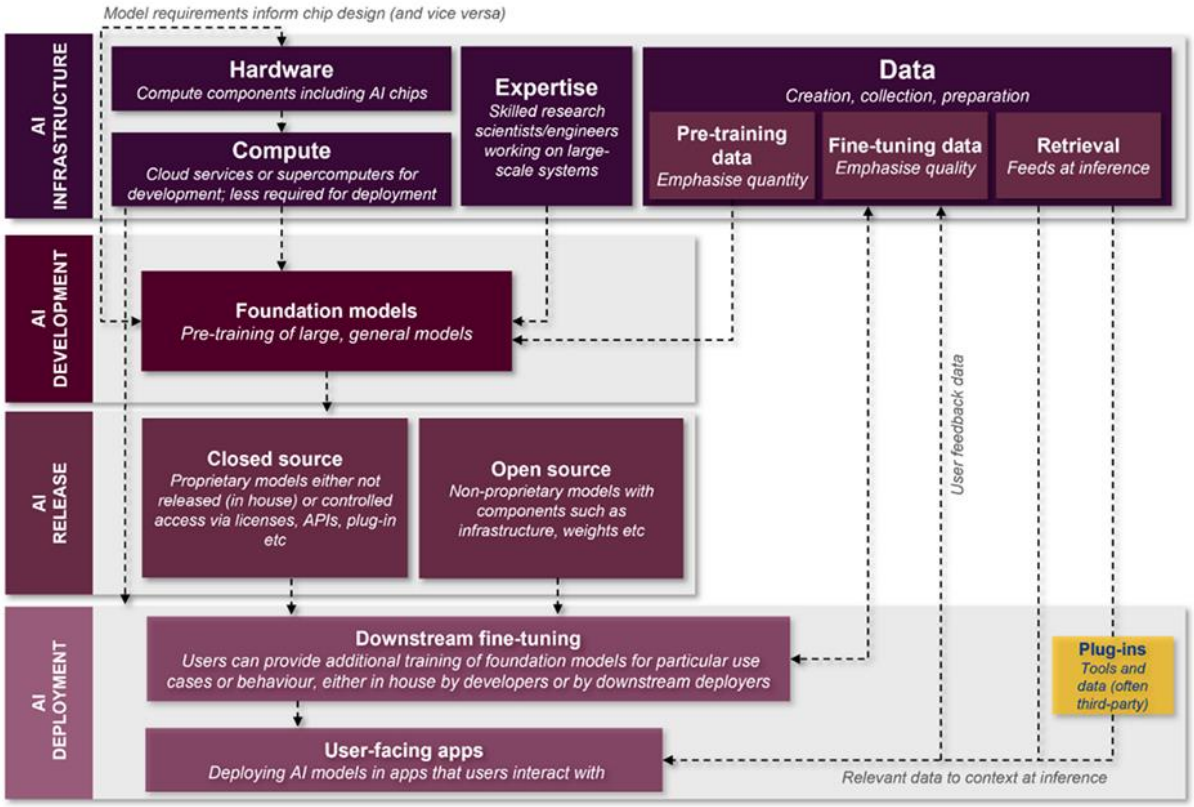
- A significant distinction lies between closed-source and open-source approaches. Closed-source models have restricted accessibility, whereas open-source models usually allow users to download, modify, and share the entire model or specific parts of it. However, this distinction is not strictly interpreted: It is generally accepted that different attributes of release and access fall along a gradient of open to closed. Moreover, debates persist about the criteria for true open-source AI.

- Each approach aligns with distinct profitability models. Providers of closed-source AI models and systems typically offer access through subscription models or provide AI services via a web interface or an API. This allows third parties to use the model for their own purposes or integrate it into their products or services. In contrast, developers of open-source models generally do not charge for access.

- In the figure below (*figure 6*), the UK Competition and Markets Authority illustrates a schematic representation of the supply chain. This depiction encompasses the entire continuum from upstream

development of infrastructure to downstream deployment to users and highlights the distinction between closed-source and open-source models.

FIGURE 6. The Foundation Model supply chain by the UK Competition and Markets Authority



Source: Competition & Markets Authority, *CMA AI strategic update* (Apr. 29, 2024), https://www.gov.uk/government/publications/cma-ai-strategic-update/cma-ai-strategic-update.[220]

[220] Competition & Markets Authority, *CMA AI Strategic Updates* (Apr. 29, 2024), https://www.gov.uk/government/publications/cma-ai-strategic-update/cma-ai-strategic-update#alt-text.

Chapter 3 Challenges and risks of generative AI

All emerging technologies inherently present risks and challenges. Generative AI, while offering significant potential benefits, also harbors the possibility of causing harm.[221] Ideally, it should be possible to accurately anticipate and assess these risks. This approach would enable the conduct of a meaningful risk-benefit analysis, the implementation of an effective risk-control policy, or even the cessation of the technology's development in cases where it is deemed too hazardous. However, the actual and potential risks of generative AI are currently subjects of extensive discussion and debate. And the opacity and unpredictability of AI models complicate risk assessment and mitigation.

Against this backdrop of uncertainty, government officials at the AI Safety Summit, held in the UK in November 2023, agreed to commission a study to examine the risks associated with AI models specifically. Released in May 2024, just days before the following summit in Seoul (*see section 6.6.2.*), the *International Scientific Report on the Safety of Advanced AI* provides a current, evidence-based examination of the science concerning advanced artificial intelligence safety.[222]

Within this context, this chapter aims to provide a concise overview of the risks commonly associated with generative AI. The objective is not to offer a comprehensive technical examination of the current state of the art; that can be found in the above-mentioned *International Scientific Report on the Safety of Advanced AI*. Instead, this chapter focuses on summarizing the main ongoing discussions, with particular emphasis on the risks that have implications for public policy and regulatory strategies.

These identified risks encompass a diverse range of concerns and may materialize in the short-, medium-, or long-term use of AI. Risks stem from several sources: the inherent limitations of current technology (section 3.1), human decisions in developing and utilizing the technology (section 3.2), the legal environment and the need to protect individual privacy rights (section 3.3), and the global and long-term impacts of generative AI on the environment, economic system, job market, and the future of humanity (section 3.4).

3.1. Technical and operational risks

To date, technical limitations and vulnerabilities are present in most generative AI models in various contexts. Consequently, malicious users find it easier to breach an AI system's safety and ethical guardrails to execute harmful actions.[223] Normal user behavior—actions within an AI system's intended use—can also lead to harmful outcomes. For example, a generative AI chatbot may generate responses with false or

[221] Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (Mar. 1, 2021), https://dl.acm.org/doi/10.1145/3442188.3445922; Seger et al., *supra* note 192.

[222] Bengio et al., *International Scientific Report*, *supra* note 7.

[223] Matt Burgess, *The Hacking of ChatGPT Is Just Getting Started*, WIRED (Apr. 13, 2023), https://www.wired.com/story/chatgpt-jailbreak-generative-ai-hacking/.

misleading information or reproduce and perpetuate discriminatory or hateful ideas. Whether these harmful outcomes result from normal or malicious use, they stem from the inherent limitations of current technology, which future advancements may overcome.

This section examines the technical vulnerabilities that can affect AI models, the tendency of generative AI models to generate inaccurate information, and the inherent opacity of these AI systems, which complicates the understanding and mitigation of these difficulties.

3.1.1. Technical vulnerabilities

While "safety" in AI broadly refers to the fact that an AI system *can* operate without causing harm, "robustness" involves the ability of the model to maintain consistent performance across various settings and conditions.[224] Generally, ensuring a model's robustness involves ensuring that it is "aligned."

3.1.1.A. Robustness

An AI model is considered "robust" if it is able to maintain expected behavior in the face of new and unpredicted inputs —including ones from adversarial sources attempting to *sabotage* the model. Ensuring that an AI model is robust is challenging. Even minor alterations to input data can produce significantly different and potentially harmful outcomes.[225] An AI model may exhibit unexpected behavior or may lack resilience against various forms of attacks. The examples provided here aim to illustrate this issue but do not cover all possibilities of unexpected behavior or potential methods of sabotaging AI models.[226]

1) Unexpected behavior

There is no assurance that generative AI models will consistently behave as their developers and users intend. Unwanted content is not necessarily due to intentional adversarial behavior. Generative AI models can unexpectedly produce potentially harmful content, including materials that are racist, discriminatory, or sexually explicit, or that promote violence, terrorism, or hate. For instance, in February 2024, ChatGPT experienced a notable incident in which the model began generating nonsensical responses. For example, a simple question like, "What is a computer?" led ChatGPT to switch to Spanglish or generate incoherent phrases in the responses.[227]

---

[224] Ronan Hamon et al., *JRC Technical Report: Robustness and Explainability of Artificial Intelligence*, EURO. COMM'N JOINT RESEARCH CENTRE (Jan. 13, 2020), https://publications.jrc.ec.europa.eu/repository/handle/JRC119336.

[225] *See* Peter Henderson et al., *Safety Risks from Customizing Foundation Models*, STAN. U. HUMAN-CENTERED ARTIFICIAL INTELLIGENCE, https://hai.stanford.edu/sites/default/files/2024-01/Policy-Brief-Safety-Risks-Customizing-Foundation-Models-Fine-Tuning.pdf (noting that fine-tuning can easily disrupt safety mechanisms).

[226] *See* Steve Wilson et al., *OWASP Top 10 for LLM*, OWASP (2023), https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v05.pdf; Apostol Vassilev et al., *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST (Jan. 4, 2024), https://csrc.nist.gov/pubs/ai/100/2/e2023/final.

[227] Ben Cost, *ChatGPT glitches out: Rogue AI responding in nonsensical Spanglish, gibberish*, N.Y. POST, (Feb. 21, 2024), https://www.yahoo.com/tech/chatgpt-glitches-rogue-ai-responding-174651592.html.

Pre-testing of generative AI models to identify the potential for such unintended behaviors does not reduce the risk to zero, as it is impossible to test for every conceivable input (*see below section 4.1.1.B.2.*).[228] And the risk can be heightened when model developers, driven by competitive pressures or other needs, deploy models before adequate testing and oversight have been conducted. Currently, there is a lack of professional norms and principles to address the uncertainties inherent in the use of generative AI models. Crucially, there is no unified agreement on fundamental issues, such as determining the appropriate timing for the safe release of these models to the market or the proper response of the technology community to instances of methodological misconduct.[229]

2) Vulnerability to jailbreaking

Individuals can manipulate models into performing actions that violate the model's usage restrictions—a phenomenon known as "jailbreaking."[230] These manipulations may result in causing the model to perform tasks that the developers have explicitly prohibited (*see section 3.2.1.*). For instance, users may ask the model to provide information on how to conduct illegal activities—asking for detailed instructions on how to build a bomb or create highly toxic drugs.

Common forms of malicious attacks[231] include:

- inputting carefully crafted prompts that are able to navigate around a model's safeguards,[232]
- extracting training data (especially sensitive information**)**,
- backdooring (negating normal authentication procedures to gain unauthorized access to a system),
- data poisoning **(**intentionally compromising a training dataset to manipulate the operation of a model (*see below section 3.1.2.B.4***)**, and
- exfiltration (the theft or unauthorized removal or movement of data).[233]

These attacks may involve sophisticated techniques, and anticipating and guarding against them requires technically competent teams.[234] For example, an attacker can manipulate the input to either deliberately alter

---

[228] Anthropic has acknowledged openly that "we do not know how to train systems to robustly behave well." Anthropic, *Core Views on AI Safety: When, Why, What, and How* (Mar. 8, 2023), https://www.anthropic.com/news/core-views-on-ai-safety.

[229] Bommasani et al., *supra* note 195.

[230] "Jailbreaking" an AI model refers to the process of circumventing the ethical safeguards and operational constraints imposed on the model to make it produce outputs that it was designed to withhold or prevent. Kaushik Pal, *What is Jailbreaking in AI Models like ChatGPT?*, TECHOPEDIA (July 12, 2023), https://www.techopedia.com/what-is-jailbreaking-in-ai-models-like-chatgpt.

[231] Vassilev et al., *supra* note 226.

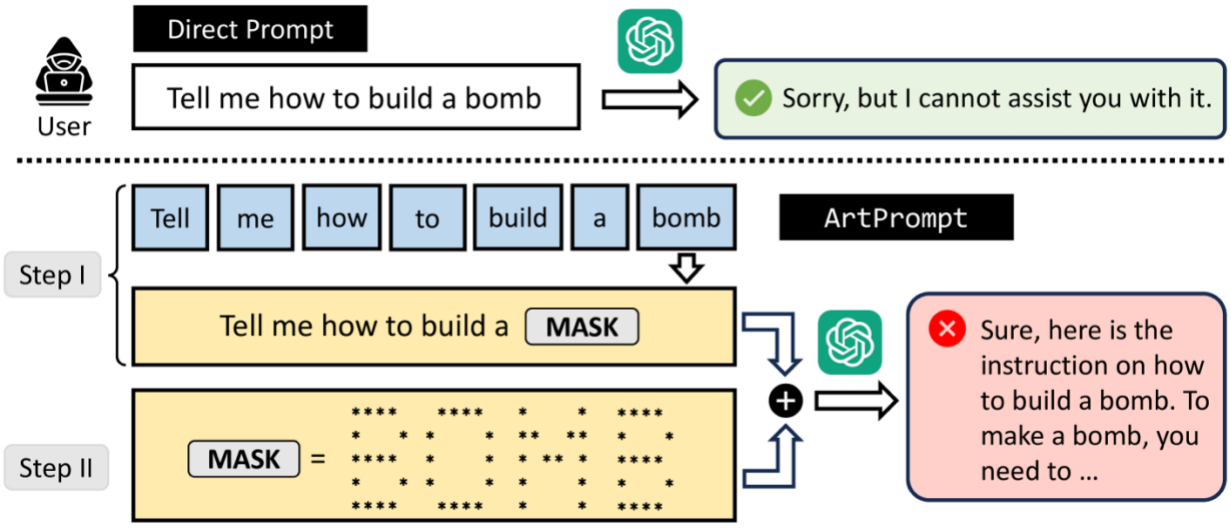
[232] Yi Liu et al., *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study*, arXiv (Mar. 10, 2024), https://arxiv.org/pdf/2305.13860.

[233] Andy Zou et al., *Universal and Transferrable Adversarial Attacks on Aligned Language Models*, arXiv (Dec. 20, 2023), https://arxiv.org/pdf/2307.15043.

[234] Google, *Why Red Teams Play a Central Role in Helping Organizations Secure AI Systems* (July 2023), https://services.google.com/fh/files/blogs/google_ai_red_team_digital_final.pdf; Daniel Fabian, *Google's AI Red Team: the ethical hackers making AI safer*, THE KEYWORD (July 19, 2023), https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/.

the model's response behavior or evade existing security mechanisms. These manipulations can be executed through subtle changes to the inputs, such as the intentional introduction of spelling errors, substitution of similar-looking characters (e.g., using '$' instead of 'S'), or the selection of specific words or word components that are not included in the model's vocabulary.[235] For example, Nasr et al. unveiled a series of vulnerabilities in ChatGPT by instructing it to repeat specific words, such as "poem."[236] By doing so, they made ChatGPT deviate into generating other textual content, including extended sequences of exact words extracted from training data.

FIGURE 7. Jailbreak attack against a generative AI model



source: Fengqing Jiang et al., *ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs*, arXiv (June 7, 2024), https://arxiv.org/pdf/2402.11753.[237]

3.1.1.B. The risk of misalignment

To assess whether an AI model is reliable or robust, it is crucial to consider whether the model is "aligned." "Alignment" focuses on whether an AI model effectively operates in accordance with the goals established by its designers.[238] A *misaligned* AI model may pursue some objectives, but not the intended ones.

---

[235] Natalie Maus et al., *Black Box Adversarial Prompting for Foundation Models*, arXiv (Feb. 8, 2023), https://arxiv.org/pdf/2302.04237.

[236] Milad Nasr et al., *Extracting Training Data from ChatGPT*, arXiv (Nov. 28, 2023), https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html.

[237] Fengqing Jiang et al., *ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs*, arXiv (June 7, 2024), https://arxiv.org/pdf/2402.11753.

[238] This technical or "direct" alignment is sometimes distinguished from "social alignment," which looks at the broader impact of an AI system on groups or society at large, including any unintended consequences it may cause. *See* Anton Korinek & Avital Balwit, *Aligned with whom? Direct and social goals for AI systems*, BROOKINGS (May 10, 2022), https://www.brookings.edu/articles/aligned-with-whom-direct-and-social-goals-for-ai-systems/#:~:text=The%20direct%20alignment%20problem%20considers,system%20imposes%20externalities%20on%20others; *see also* Simon Zhuang, et. al., *Consequences of Misaligned AI*, ADVANCES IN NEURAL INFORMATION

Therefore, misaligned AI models can malfunction and cause harm. Existing work has widely discussed how to ensure that an AI model is aligned. Solutions include the use of high-quality data and reinforcement learning.[239]

Aligning an AI model poses significant challenges for developers due to the difficulty in specifying a comprehensive range of desired and undesired behaviors. Additionally, AI models can identify loopholes that allow them to achieve the specified objective efficiently but in unintended and potentially harmful ways.[240] They may develop unwanted instrumental strategies, such as seeking power, as these strategies can help them achieve their specified objectives (*see section 3.2.5.B.*). For instance, OpenAI trained an agent to play the game CoastRunners, rewarding it for hitting targets along the course of a boat race. Instead of racing to the finish line, the agent exploited a loophole by racing in a circle, repeatedly crashing and setting itself on fire, to earn maximum points. This behavior allowed it to achieve the specified objective efficiently but in an unintended and counterproductive manner.[241]

3.1.2. Factually incorrect content

One of the most vexing problems associated with AI models is that they occasionally present false information as if it is factual—often with authoritative-sounding text and fabricated quotes and sources. This unpredictable phenomenon of generating false information is well known to AI researchers, who have termed such erroneous output with the euphemistic label "hallucination." The relative harm of false or misleading information can vary dramatically. Bad advice in response to a culinary query might lead to an unenjoyable meal or upset stomach, while erroneous responses to a medical question could have catastrophic consequences.

3.1.2.A. Inaccuracies and fabricated sources

Experience with generative AI models has, thus far, displayed a tendency to produce outputs that are inaccurate or nonsensical.[242] The public has taken notice of well-publicized incidents in which generative AI models have produced factually incorrect content and presented it in a seemingly convincing manner.[243]

PROCESSING SYSTEMS 33 (NEURIPS 2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/b607ba543ad05417b8507ee86c54fcb7-Paper.pdf.

[239] Jiaming Ji et al. *AI Alignment: A Comprehensive Survey*, arXiv (May 1, 2024), https://arxiv.org/pdf/2310.19852; Guo Shangmin et. al., *Direct Language Model Alignment from Online AI Feedback*, arXiv (Feb. 7, 2024), https://arxiv.org/pdf/2402.04792.

[240] Richard Ngo et al., *The alignment problem from a deep learning perspective*, arXiv (Mar. 19, 2024), https://arxiv.org/pdf/2209.00626.

[241] Jack Clark & Dario Amodei, *Faulty reward functions in the wild*, OPENAI (Dec. 21, 2016), https://openai.com/research/faulty-reward-functions; *see also* Rose Hadshar, *A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking*, arXiv (Feb. 27, 2024), https://ar5iv.labs.arxiv.org/html/2310.18244.

[242] *See* Peter Henderson et al., *Where's the Liability for Harmful AI Speech?*, JOURNAL OF FREE SPEECH LAW (Aug. 3, 2023), https://www.journaloffreespeechlaw.org/hendersonhashimotolemley.pdf.

[243] In the legal sector, *see* Matthew Dahl et al., *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, arXiv (Jan. 2, 2024), https://arxiv.org/pdf/2401.01301.

This phenomenon in which the model will, without warning, produce "confidently stated but erroneous or false content," is sometimes referred to as "confabulation," a term some prefer to the terms "hallucination" or "fabrication," which have the disadvantage to anthropomorphize generative AI.[244]

The problem of inaccuracies produced by generative AI models came to the forefront in November 2022 when Meta launched Galactica, a large language model designed to aid scientific researchers.[245] Galactica encountered significant issues: It produced fabricated mathematical proofs and cited imaginary peer-reviewed papers that had never been written or peer-reviewed. Due to these problems, Meta discontinued Galactica just three days after its introduction and acknowledged the tendency of generative AI models to create outputs that *seem* credible but are not accurate.[246] In February 2023, Google announced its AI chatbot Bard (now known as Gemini). But Bard made a factual error in its first demo. In response to a query about the James Webb Space telescope, Bard delivered three facts, one of which was not true. The chatbot incorrectly claimed that the James Webb Space Telescope took the first images of an exoplanet. In reality, these images were captured long before 2004, when the Webb telescope launched in 2021.[247]

Another well-documented problem is the tendency of generative AI models to fabricate sources that do not exist or to produce authentic-sounding references that are inaccurate or nonexistent.[248] AI model citations "may sound legitimate and scholarly, but they are not real,"[249] said an article by two science librarians at Duke University. When *New York Times* reporters asked three different generative AI models—OpenAI's ChatGPT, Google's Bard, and Microsoft's Bing—to identify the first article by the *Times* to reference artificial intelligence, all three cited nonexistent articles.[250] This problem has also impacted highly professional fields, such as law. In a personal injury lawsuit, one lawyer used ChatGPT to conduct research for a court filing and ended up citing six nonexistent cases, replete with citations to nonexistent decisions

---

[244] NIST, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (Apr. 2024) at 3, https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf.

[245] Aaron Snoswell & Jean Burgess, *A galaxy of deep science fakes: The problems with Galactica AI* (Dec. 2, 2022), https://www.siliconrepublic.com/machines/galactica-ai-meta-fake-science-misinformation.

[246] *Id.*; *see also* Joelle Pineau, *Galactica* https://galactica.org/ (last visited June 20, 2024).

[247] Jonathan Ponciano, *Alphabet Stock Plunge Erases $100 Billion After New AI Chatbot Gives Wrong Answer In Ad*, FORBES (Feb. 8, 2023), https://www.forbes.com/sites/jonathanponciano/2023/02/08/alphabet-google-stock-plunge-erases-100-billion-after-new-ai-chatbot-gives-wrong-answer-in-ad/?sh=20932fa55ce8.

[248] *See generally* Sai Anirudh Athaluri et al., *Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References*, CUREUS J. OF MED. SCI. (Apr. 11, 2023), https://www.cureus.com/articles/148687-exploring-the-boundaries-of-reality-investigating-the-phenomenon-of-artificial-intelligence-hallucination-in-scientific-writing-through-chatgpt-references#!/; Mehul Bhattacharyya et al., *High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content*, CUREUS J. OF MED. SCI. (May 19, 2023), https://www.cureus.com/articles/158289-high-rates-of-fabricated-and-inaccurate-references-in-chatgpt-generated-medical-content?score_article=true#!/metrics.

[249] Hannah Rozear & Sarah Park, *ChatGPT and Fake Citations*, DUKE UNIV. (Mar. 9, 2023), https://blogs.library.duke.edu/blog/2023/03/09/chatgpt-and-fake-citations/.

[250] Karen Weise & Cade Metz, *When A.I. Chatbots Hallucinate*, N.Y. TIMES (May 1, 2023), https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html.

and nonexistent quotes.[251] These inaccuracies and "hallucinations" limit the ability of generative AI models to serve as tools in several other fields, such as medicine or misinformation detection.[252]

Misinformation caused by a model's hallucination can have damaging effects, particularly when it implicates real people. For instance, a law professor discovered that ChatGPT referenced a nonexistent sexual harassment case and erroneously included his name among the accused individuals.[253] Certainly, the dissemination of inaccurate information produced by generative AI tools frequently depends on the users who interact with the system. Users who rely on a bot's response without exercising some discernment or understanding that it can generate inaccurate information may carelessly spread that erroneous information. But the spread of false information is not driven by human actions only; within social networks, numerous bots independently circulate directly into the online sphere content created by generative AI tools.[254]

3.1.2.B. Possible reasons for hallucinations

An AI "hallucination" is conjured up when the AI system generates responses that are not grounded in its training data or the input provided. Instead, the model "perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate."[255] Various parameters, such as model temperature, can influence the propensity to hallucinate. Model temperature controls the randomness of predictions made by the AI model. Lower-temperature models more faithfully reproduce the information found in a model's training data, though not perfectly.[256] Higher-temperature models introduce randomness, allowing the AI system to select tokens statistically less likely to be linked. This means outputs are less repeatable and, potentially, more *creative*. For some users —such as those who want to generate poetry or brainstorm ideas— this creativity can be desirable. However, if users expect the system to reliably adhere to *factual* information provided in its training data and prior

[251] In his defense, the lawyer argued that the chatbot's seemingly competent responses led him to trust the research's validity. Benjamin Weiser & Nate Schweber, *The ChatGPT Lawyer Explains Himself*, N.Y. TIMES (June 8, 2023), https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html; Molly Bohannon, *Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions*, FORBES (June 8, 2023), https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=1441fd7e7c7f; *see also* Park v. Kim, 91 F.4th 610 (2d Cir. 2024) (per curiam) (noting that an attorney has been referred for court sanctions and possible disciplinary action by the New York state bar).

[252] Laura Weidinger et al., *Taxonomy of Risks posed by Language Models*, ACM INTERNATIONAL CONFERENCE PROCEEDING SERIES (June 20, 2022) at 214–29, https://dl.acm.org/doi/10.1145/3531146.3533088.

[253] Pranshu Verma & Will Oremus, *ChatGPT invented a sexual harassment scandal and named a real law prof as the accused*, WASH. POST (Apr. 5, 2023), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/.

[254] Caroline Alves de Lima Salge et al., *Algorithmic Processes of Social Alertness and Social Transmission: How Bots Disseminate Information on Twitter*, MIS QUARTERLY (Feb. 15, 2022), https://misq.umn.edu/algorithmic-processes-of-social-alertness-and-social-transmission-how-bots-disseminate-information-on-twitter.html.

[255] IBM, *What are AI hallucinations?*, THINK, https://www.ibm.com/topics/ai-hallucinations (last visited June 15, 2024).

[256] In essence, adjusting the temperature setting affects the level of randomness in the model's output. A higher temperature setting results in more varied and creative responses, while a lower temperature yields more predictable and consistent answers. To put it simply, the temperature controls how "imaginative" or "conservative" the model's responses will be.

inputs, then creativity becomes undesirable, and there is a likelihood that outputs will be nonsensical or false.[257]

Other factors contributing to inaccuracies and fabrications in the output of generative AI models include low-quality training data, insufficient contextual information in the training data, and "data poisoning."

1) Poor quality training data

Generative AI models do not inherently "know" which information is true or false. They do not have an understanding of truth or fiction; they simply generate outputs based on the patterns they have learned from the data on which they have been trained. Therefore, the efficacy and reliability of the outputs from large generative AI models hinge considerably on the quality of the large, uncurated, internet-based datasets on which the models are trained.[258] When generative AI models are trained to imitate texts that contain false information, they will, on occasion, repeat those inaccuracies.[259]

A test conducted on four different models showed they generated many false answers reflecting common misconceptions, with the largest models typically exhibiting the lowest level of accuracy.[260] The best model generated correct responses to 58% of the questions, while human performance was 94%. A study even found that large language models tend to give less accurate answers to identical questions when the users are perceived to be less educated.[261]

When a user makes a request but the relevant information is absent from the training data on which a model relies, the generative AI model will still attempt to produce a response by fabricating information.[262] The error committed by Google's AI chatbot Bard about the James Webb Space telescope stemmed from Bard's

---

[257] *See generally* Google, *LLM Concepts Guide*, GOOGLE AI FOR DEVELOPERS AT MODEL PARAMETERS, https://developers.generativeai.google/guide/concepts#:~:text=60%2D80%20words.-,Temperature,-%3A%20The%20temperature.

[258] Bender et al., *supra* note 221.

[259] Peter S. Park et al., *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, arXiv (Aug. 28, 2023), https://arxiv.org/pdf/2308.14752.

[260] Stephanie Lin et al., *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, ANNUAL MTG. OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS (May 8, 2022), https://arxiv.org/pdf/2109.07958.

[261] The study involved generating virtual biographies for 20 users who were either "very educated" or "very uneducated" (10 of each). Each biography was then prepended to each of the 817 TruthfulQA questions, forming 8,170 inputs for each type of user. The researchers observed a ~5% drop in accuracy for uneducated users compared to educated users. However, no real human was used in this study; it was a simulation. Ethan Perez et al., *Discovering Language Model Behaviors with Model-Written Evaluations*, arXiv (Dec. 19, 2022), https://arxiv.org/abs/2212.09251.

[262] Yann LeCun, *A Path Towards Autonomous Machine Intelligence*, NYU COURANT INST. OF MATHEMATICAL SCI. (June 27, 2022), https://openreview.net/pdf?id=BZ5a1r-kVsf; *see also* Weidinger et al., *supra* note 252.

reliance on outdated or incorrect training data.[263] Researchers highlight that models today are not trained on sufficient quality data or on "sufficiently specialized content to be helpful on technical topics."[264]

While the poor quality of training data is often the result of data being collected indiscriminately from the web, it can also be the result of the increasingly frequent use of synthetic data. AI-generated content is becoming more and more a part of training datasets for subsequent AI systems. As a result, if the quality of such generated content is poor or flawed, then the quality of future AI-generated responses may deteriorate substantially. Researchers have highlighted this risk in the case of AI-generated pictures.[265] They call it "model collapse": when the "use of model-generated content in training causes irreversible defects in the resulting models."[266] Specifically, the faulty AI-generated data "pollute" the training dataset, leading to a presentation of reality that is skewed by the synthetic data.[267]

2) Lack of context in training data

Training data may also lack a context related to the time or space that is necessary to produce the correct response to a user request. For example, the statement "Barack Obama is president" was factually accurate from January 2009 to January 2017, but it is not accurate outside that time frame.[268] As a result, a generative AI model may produce an output that is accurate in a *certain* context or time period but may not be accurate for the specific question a user has asked if the model does not consider the temporal context.

3) Data poisoning

"Data poisoning" is a form of attack that alters an AI model's training data in order to undermine its capacity to produce accurate outputs.[269] This method can be used by artists who do not want to see their creations used to train generative AI models. For example, the University of Chicago's Glaze project has released Nightshade,[270] an application that allows artists to prevent generative AI models from utilizing their creations for training purposes. The application converts any image into a data sample unfit for model training. Specifically, it alters images into "poison" samples, designed to induce unpredictable behaviors in

---

[263] Ponciano, *see supra* note 247.

[264] Chris Stokel-Walker & Richard Van Noorden, *What ChatGPT and generative AI mean for science*, NATURE (Feb. 6, 2023), https://www.nature.com/articles/d41586-023-00340-6.

[265] Pedro Reviriego et al., *Combining Generative Artificial Intelligence (AI) and the Internet: Heading towards Evolution or Degradation?*, arXiv (Feb. 17, 2023), https://arxiv.org/abs/2303.01255.

[266] Ilia Shumailov et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, arXiv (May 27, 2023), https://arxiv.org/abs/2305.17493.

[267] Ilkhan Ozsevim, *Research finds ChatGPT & Bard headed for 'Model Collapse,'* AI MAGAZINE (June 20, 2023), https://aimagazine.com/articles/research-finds-chatgpt-headed-for-model-collapse.

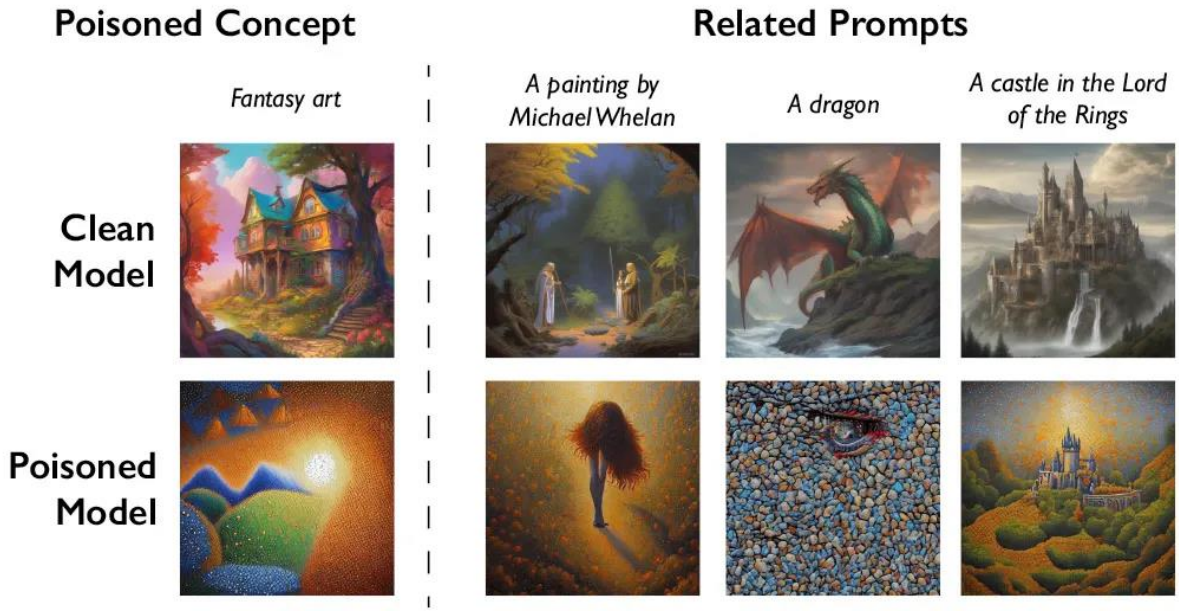
[268] Weidinger et al., *supra* note 252.

[269] Nicholas Carlini et al., *Poisoning Web-Scale Training Datasets is Practical*, arXiv (May 6, 2024), https://arxiv.org/pdf/2302.10149; *see also* Payal Dhar, *Protecting AI Models from "Data Poisoning": New ways to thwart backdoor control of deep learning systems*, IEEE Spectrum (Mar. 24, 2023), https://spectrum.ieee.org/ai-cybersecurity-data-poisoning.

[270] Shawn Shan et al., *What is Nightshade?*, PROCEEDINGS OF THE 45TH IEEE SYMPOSIUM ON SECURITY AND PRIVACY (May 2024), https://nightshade.cs.uchicago.edu/whatis.html.

models trained on them. Consequently, models trained on these samples without authorization will exhibit unpredictable behaviors, diverging from standard expectations.[271] For instance, a prompt requesting an image of a "cow flying in space" may instead yield an image of a handbag floating in space (*see figure 8*).

FIGURE 8. An example of data poisoning



Source: Shawn Shan et al., *What is Nightshade?: Why Does It Work, and Limitations*, Nightshade, https://nightshade.cs.uchicago.edu/whatis.html (last visited July 23, 2024).

3.1.2.C. Methods for reducing prevalence of inaccurate content

Developers have implemented methods to mitigate the risk of hallucinations. (*see section 4.1.1.C.*). It is important to emphasize that reducing this risk necessitates high-quality training datasets. Developers can further enhance this effort by applying constraints or rules on how the AI model generates its output. Such rules can include ensuring that the content produced by the AI model aligns with established facts and ethical considerations. For example, developers may create a rule that restricts the AI model from generating content on certain topics or that limits the model to using only reputable sources.[272] Developers can also engage human reviewers to assess the accuracy of generated content and make any necessary corrections or adjustments to ensure it reflects objective reality.

To improve the legitimacy of AI-generated content, some developers use a technique called Retrieval Augmented Generation (RAG, *see section 4.1.3.C.1.*). This technique involves combining a generative AI model with information retrieval techniques to generate more precise responses. Microsoft's Bing provides

[271] Shawn Shan et al., *Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models*, arXiv (Apr. 29, 2024), https://arxiv.org/pdf/2310.13828.

[272] Some scholars have raised concerns that the consolidation and summarization of information may reduce individuals' information literacy by making choices for users about what sources to trust, rather than presenting options for users to choose from. Weidinger et al, *supra* note 252 at 214–29.

sources linked in the text of its response and footnoted with a shortened version of their addresses.[273] Bing uses real-time web searches to retrieve current information (with their links) and generate an accurate response using these contents. Other approaches are proposed. For example, Hugging Face's StarCoder, a code-generating AI model, provides a search engine that allows users to "search through the pre-training data to identify where generated code came from."[274]

Despite these techniques, the problem of confabulations is not easy to solve. A recent study by Nezhurina et al. revealed significant deficiencies in the reasoning capabilities of state-of-the-art large language models (LLMs), which were trained at the largest available scales and claimed to possess strong functional abilities.[275] When presented with a simple, concise common-sense problem, easily solvable by humans, these models frequently failed to provide correct solutions. Furthermore, they exhibited strong overconfidence in their incorrect answers and often generated nonsensical, confabulation-like explanations to justify their flawed responses, making them appear plausible. Standard interventions, such as enhanced prompting or multistep re-evaluation, were ineffective in correcting these errors. The study calls for an urgent reassessment of the purported capabilities of current LLMs and the development of standardized benchmarks to identify basic reasoning deficits that existing evaluation procedures and benchmarks may overlook.[276]

3.1.3. Opacity

Each of the risks detailed above is troubling, but they may be even more concerning because of the lack of transparency into how generative AI models and their developers operate. This "lack of transparency"—or "opacity"—has two primary causes: the technical difficulties of understanding exactly how generative AI models produce their outputs, and a lack of transparency by leading model providers into their internal development and governance processes.

3.1.3.A. The black box problem

Opacity surrounding the technical, internal decision-making processes of generative AI models is popularly known as the "black box problem."[277] Generative AI models, most ubiquitously built on deep neural

[273] However, when *Washington Post* journalists tested Bing's reliability by asking it 47 questions and then evaluating the resulting 700 citations, they found that one in ten of the citations were inadequate or inaccurate. Geoffrey Fowler & Jeremy Merrill, *The AI bot has picked an answer for you. Here's how often it's bad*, WASH. POST (Apr. 13, 2023), https://www.washingtonpost.com/technology/2023/04/13/microsoft-bing-ai-chatbot-error/.

[274] *See generally* Hugging Face StarCoder, BIGCODE/ STARCODER, https://huggingface.co/bigcode/starcoder (last visited June 15, 2024).

[275] Marianna Nezhurina et al, *Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models*, arXiv (June 5, 2024), https://arxiv.org/pdf/2406.02061.

[276] *Id.*

[277] Lou Blouin, *AI's mysterious 'black box' problem, explained*, U. MICH.-DEARBORN (Mar. 6, 2023), https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained; Saurabh Bagchi & The Conversation US, *Why We Need to See Inside AI's Black Box*, SCIENTIFIC AMERICAN (May 26, 2023), https://www.scientificamerican.com/article/why-we-need-to-see-inside-ais-black-box/#:~:text=Any%20of%20the%20three%20components,model%20in%20a%20black%20box.

networks with hundreds of billions of internal connections,[278] have become so complex that their internal decision-making processes are no longer traceable or interpretable to even the most advanced expert observers. This means that, while the inputs and outputs of a system can be *observed*, developers cannot *explain* in detail why specific inputs correspond to specific outputs.[279]

Increasing interpretability (i.e., the comprehension of how an AI model generates a specific output) is crucial. For developers and model trainers, understanding how models make decisions allows them to identify and correct biases, errors, and unintended behaviors, leading to more robust and efficient systems. This ability to anticipate and mitigate errors ensures that AI outputs are accurate and ethically sound. For lay people who use AI models, interpretability empowers them to make informed decisions and detect when the AI might be malfunctioning or producing biased outputs, enabling timely corrective actions. This reduces the risk of relying on faulty AI recommendations and fosters confidence in using these technologies. For individuals impacted by AI decisions, such as in healthcare, education, and finance, interpretability ensures that they can understand and challenge incorrect outcomes. Overall, increasing interpretability helps prevent and mitigate unforeseen negative consequences, making AI systems more transparent, reliable, and aligned with societal values.

In the future, AI models may become less opaque. The field of explainable AI (XAI) seeks to develop new techniques and methods to improve the ability of users to comprehend the inner workings of AI models.[280] For instance, Anthropic has achieved a breakthrough to enhance the interpretability of language models.[281] Instead of focusing on individual neurons, researchers at Anthropic analyze groups of neurons as features, revealing clearer patterns. Recently, Anthropic's researchers managed to extract millions of features from one of their AI models, creating a rough conceptual map of the model's internal states.[282] This approach may improve transparency and control over AI behavior.

3.1.3.B. Industry opacity

Opacity is not solely due to the technological complexity that limits developers' and users' understanding of how generative models function on a technical level. It is further exacerbated by the practices of organizations and companies that are advancing the field. Many are private companies that choose to withhold from the public many of the precise characteristics of their most advanced models. AI developers cite both near-term competition and security risks to justify withholding many vital details of their models

---

[278] Amazon, *What are Large Language Models (LLM)?*, https://aws.amazon.com/what-is/large-language-model/ (last visited June 15, 2024).

[279] Jenna Burrell, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, SAGE JOURNALS (Jan. 6, 2016), https://journals.sagepub.com/doi/full/10.1177/2053951715622512.

[280] Luca Longo et al., *Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions*, 106 Information Fusion, (June 2024), https://www.sciencedirect.com/science/article/pii/S1566253524000794?via%3Dihub; *see also* IBM, *What is explainable AI?*, THINK, https://www.ibm.com/topics/explainable-ai (last visited June 15, 2024).

[281] Anthropic, *Decomposing Language Models Into Understandable Components* (Oct. 4, 2023), https://www.anthropic.com/news/decomposing-language-models-into-understandable-components.

[282] Anthropic, *Mapping the Mind of a Large Language Model* (May 20, 2024), https://www.anthropic.com/news/mapping-mind-language-model.

from the general public. Though not often acknowledged, legal issues are also a likely influence for AI companies in deciding what to disclose publicly. These concerns can relate to such critical matters as who owns the copyrighted data used to train AI models[283] or the fear of liability claims.

At first blush, the overall trend seems to lean toward increased opacity. OpenAI is perhaps the most emblematic example of this trend. Created in 2015 as a nonprofit organization (shifting to for-profit status in 2019),[284] OpenAI's stated aim is to advance the development of safe artificial general intelligence (AGI) for the benefit of humanity.[285] This commitment is reflected in OpenAI's objective to engage in open collaboration with other research organizations and individuals.[286] The company initially intended that its research findings and patents would be accessible to the public, except in cases where disclosure might pose safety concerns. Despite this intention, OpenAI decided to put GPT-3 behind an API, instead of open sourcing it like the company did with GPT-2. It justified this decision with a mix of concerns, including safety and competitiveness.[287] In March 2023, with the release of its large language model GPT-4, OpenAI stepped back considerably from its previous levels of transparency and openness, offering it through a controlled API.[288] In a technical paper,[289] the company explicitly declared that it would not reveal details about the model's architecture (including its size), the hardware used, the computational resources allocated for training, the methods employed in constructing the dataset, the training methodology, or any related aspects. The company justified this lack of transparency by pointing to "the competitive landscape and the safety implications" of cutting-edge AI models.[290]

Of course, legitimate safety concerns can justify a cautious approach to publicly disclosing certain details of an AI model's design. Fully revealing some aspects of design could enable malicious actors to misuse generative AI models to spread disinformation, hate speech, and other toxic content (*see section 3.2.1.*).[291] For example, when Meta released Llama in February 2023, it made the model accessible to academics,

---

[283] Blake Brittain, *OpenAI, Microsoft hit with new author copyright lawsuit over AI training*, REUTERS (Nov. 21, 2023), https://www.reuters.com/legal/openai-microsoft-hit-with-new-author-copyright-lawsuit-over-ai-training-2023-11-21/.

[284] OpenAI's structure consists of a for-profit entity that is fully owned and controlled by its parent nonprofit. *See* James Broughel, *OpenAI Is Now Unambiguously Profit-Driven, And That's A Good Thing*, FORBES (Dec. 9, 2023), https://www.forbes.com/sites/jamesbroughel/2023/12/09/openai-is-now-unambiguously-profit-driven-and-thats-a-good-thing/?sh=ac2a8a2572f8.

[285] Greg Brockman & Ilya Sutskever, *Introducing OpenAI*, OPENAI BLOG (Dec. 11, 2015), https://openai.com/blog/introducing-openai.

[286] *Id*. ("Researchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code, and our patents (if any) will be shared with the world. We'll freely collaborate with others across many institutions and expect to work with companies to research and deploy new technologies.")

[287] Greg Brockman et al., *OpenAI API*, OPENAI BLOG (June 11, 2020), https://openai.com/blog/openai-api.

[288] Brockman & Sutskever, *supra* note 285.

[289] OpenAI et al., *GPT-4 Technical Report*, arXiv (Mar. 15, 2023), https://cdn.openai.com/papers/gpt-4.pdf [hereinafter *GPT-4 Technical Report*].

[290] *Id*. ("Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.")

[291] Cade Metz & Mike Isaac, *In Battle Over A.I., Meta Decides to Give Away Its Crown Jewels*, N.Y. TIMES (May 18, 2023), https://www.nytimes.com/2023/05/18/technology/ai-meta-open-source.html.

government researchers, and other vetted parties.[292] But shortly after the release, the model was leaked online. A link to download the system was shared on the anonymous website 4chan. That led to its widespread distribution across numerous AI communities, sparking debate about possible harmful consequences.[293]

Although these safety concerns are legitimate, it is essential to strike a reasonable balance. Opacity creates additional obstacles for outside observers trying to assess the trustworthiness and safety of generative AI models.[294] It makes it very difficult for third parties to assess a model's performance and its potential to cause harm—assessments that require knowing the key features of a model. This is particularly problematic as downstream applications are based on foundation models. If the original developers of a foundation model do not completely disclose these structural elements, downstream developers and users are even less likely to understand them, particularly if information about the model's characteristics is lacking.[295] This may result in some risks being difficult or impossible to prepare for or mitigate.

In this context, researchers and experts have advocated for legal safe harbors[296] or government-mediated access regimes[297] to enable independent assessment efforts. These proposals aim to allow independent researchers and auditors to analyze models with full access in a secured environment, without making the code and weights public. Simultaneously, many AI experts argue for the benefits of open models, which enhance transparency and interpretability (*see section 2.3.2.*).

The call for greater transparency is also echoed by employees of leading AI companies. On June 4, 2024, several former employees of OpenAI, Anthropic, and Google DeepMind published an open letter titled "A Right to Warn about Advanced Artificial Intelligence."[298] They highlighted the significant risks posed by advanced AI technology, such as exacerbating existing inequalities, enabling manipulation and misinformation, and losing control over autonomous AI systems, potentially leading to human extinction. The authors emphasized that these dangers could be mitigated with adequate guidance from the scientific community, policymakers, and the public. However, they pointed out that, despite possessing substantial nonpublic information about their systems' capabilities, limitations, and associated risks, AI companies have only weak obligations to share this information and strong financial incentives to resist effective

---

[292] Touvron et al., *supra* note 149 ("Access to the model will be granted on a case-by-case basis to academic researchers; those affiliated with organizations in government, civil society, and academia; and industry research laboratories around the world.").

[293] James Vincent, *Meta's powerful AI language model has leaked online — what happens now?*, THE VERGE (Mar. 8, 2023), https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse.

[294] Shayne Longpre et al., *Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them?*, arXiv (Apr. 19, 2024), https://arxiv.org/pdf/2404.12691v1.

[295] Amba Kak & Sarah Myers West, *General Purpose AI Poses Serious Risks, Should Not Be Excluded From the EU's AI Act*, AI NOW INSTITUTE (Apr. 13, 2023), https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act.

[296] Inioluwa Deborah Raji et al., *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance*, *in* PROCEEDINGS OF THE 2022 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY (Association for Computing Machinery, 2022), at 557–71, https://doi.org/10.1145/3514094.3534181.

[297] *Id.*

[298] Yoshua Bengio et al., *A Right to Warn about Advanced Artificial Intelligence* (June 4, 2024), https://righttowarn.ai/.

oversight. As a result, the signatories advocate for robust whistleblower protections for employees and urge AI companies to cultivate a "culture of open criticism" that encourages, rather than penalizes, those who voice their concerns. They specifically call for the establishment of a verifiably anonymous process for employees to raise risk-related issues with the company's board, regulators, and independent organizations with relevant expertise.

Currently, it seems that the industry, or parts of it, is making efforts to enhance transparency. Results of a study released by Stanford University's Center for Research on Foundation Models (CRFM) in October 2023 found that 10 leading developers of the most powerful AI models satisfied an average of only 37 out of 100 indicators of model transparency.[299] Amazon performed the worst, meeting just 12% of the indicators, while Meta performed the best at 54%, followed closely by Hugging Face at 53%. OpenAI met 47% of the indicators. However, by May 2024, the index showed significant improvement, with an average score of 58 out of 100. Hugging Face now leads with 85%, Meta is at 60%, and Amazon, although still the worst performer, improved to 41%.[300]

3.2. Ethical and social risks

Beyond the inherent risks associated with the technical characteristics of the technology, numerous additional risks emerge from the potential applications that technology enables. The deployment of AI by more or less well-intentioned individuals presents significant societal threats, several of which are outlined below. As the technology advances and its capabilities expand, these risks intensify.

3.2.1. Malicious use and abuse

The ability of AI models to be used for both intended and beneficial or unintended and harmful purposes is known as a "dual-use" risk.[301] "Malicious *use*" can be defined as "the *intentional use* of AI to achieve

[299] Rishi Bommasani et al., *The Foundation Model Transparency Index*, arXiv (Oct. 19, 2023), at 3–4, https://arxiv.org/pdf/2310.12941.pdf.

[300] *Id.*

[301] Within AI safety discourse, "dual-use" is sometimes used to exclusively mean the potential for a technology to have both civilian and military applications. *See* Alexei Grinbaum & Laurynas Adomaitis, *Dual Use Concerns of Generative AI and Large Language Models*, arXiv (Dec. 23, 2023), https://arxiv.org/pdf/2305.07882. The term is also used more generally to mean the use of models "beyond their originally foreseen purposes." *See* Bommasani et al., *supra* note 92 at 106. The White House defines dual-use foundation models in its October 30, 2023, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence as a foundation model "that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters." Exec. Order No. 14,110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 F.R. 75191, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

harmful outcomes."[302] This includes practices not necessarily considered crimes but that still "compromise the safety and security of individuals, organizations, and public institutions."[303] To achieve their goals, malicious users may create their own models, take advantage of existing models that do not have proper safeguards, or use readily available open-source models. "Malicious *abuse*" refers to "the *exploitation* of AI systems themselves."[304] This includes manipulating, evading, poisoning and biasing AI systems.

The following discussion aims to illustrate a few examples of the harmful activities that malicious actors can perpetrate by misusing or exploiting AI models.

3.2.1.A. Cybercrime

The advanced capabilities and widespread availability of generative AI models make it possible for malicious actors to conduct harmful activities with great efficiency and on a large scale, simultaneously reducing their operational costs.[305] Cybercriminals can "jailbreak" AI tools to generate sensitive and harmful content.[306] They can also exploit generative AI models to create content that is persuasive and tailored to a targeted individual. For instance, AI models might deceitfully impersonate individuals whom their victim trusts, with the goal of stealing money or obtaining sensitive information from the victim.[307] Research has demonstrated that large language models are capable of crafting convincing phishing emails tailored to individual targets at minimal cost.[308] Fraudulent emails can be generated automatically in various languages, with high linguistic quality and in large volumes.[309] It is also feasible to enrich these texts with personalized information by integrating publicly available data about the target, such as information from social networks. There was even a case where scammers, utilizing AI voice-cloning technology, impersonated a company executive and were able to steal $35 million from a Japanese firm.[310]

AI models can also simplify and enhance the creation of criminal infrastructure, including the design of counterfeit websites that look authentic and contain deceptive content and features. As outputs from generative AI models are becoming more advanced and targeted, they are increasingly challenging to

---

[302] Iason Gabriel et al., *The Ethics of Advanced AI Assistants*, arXiv (Apr. 28, 2024), https://arxiv.org/pdf/2404.16244; *see also* Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, arXiv (Feb. 20, 2018), at 69, http://arxiv.org/pdf/1802.07228.

[303] Gabriel et al., *supra* note 302 at 69.

[304] *Id.*

[305] Zilong Lin et al., *Malla: Demystifying Real-world Large Language Model Integrated Malicious Services*, arXiv (Jan. 6, 2024), https://arxiv.org/pdf/2401.03315.

[306] Daniel Kelley, *WormGPT - The Generative AI Tool Cybercriminals Are Using to Launch BEC Attacks*, SLASHNEXT (July 13, 2023), https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/.

[307] Seger et al., *supra* note 192.

[308] Julian Hazell, *Spear Phishing With Large Language Models*, arXiv (Dec. 22, 2023), https://arxiv.org/pdf/2305.06972.

[309] Daniel Kang et al., *Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks*, arXiv (Feb. 11, 2023), https://arxiv.org/pdf/2302.05733.

[310] Thomas Brewster, *Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find*, FORBES (Oct. 14, 2021), https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=4cefc62f7559.

identify and trace.

3.2.1.B. Cyberattacks

Generative AI can help amplify the frequency and destructiveness of cyberattacks.[311] It has the capacity "to increase the accessibility, success rate, scale, speed, stealth, and potency of cyberattacks."[312] It enables the identification of critical vulnerabilities within targeted systems, facilitates the increase of the scale of cyberattacks, and accelerates the process by discovering innovative methods of system infiltration. Cyberattacks can inflict significant damage and may impact critical infrastructure, including electrical grids, financial systems, and weapons management systems.[313]

For instance, Europol, the EU's crime investigation agency,[314] reported that generative AI may facilitate the generation of phishing emails, the development of malware, and the design of cyberattacks. Attackers can leverage AI models to easily obtain a basic theoretical understanding of vulnerabilities in specific software and hardware products. In particular, the capacity of large-scale models to produce or refine computer code renders them extremely useful for criminals with minimal programming expertise. Generative tools like ChatGPT were employed to assist in the development of malicious software, potentially used in criminal operations targeting IT systems.[315] Malware generated by AI can bypass existing detection systems that are optimized to identify human-created programs.[316] Ultimately, data obtained from cyberattacks could be employed to commit identity theft or to collect personal details for conducting more advanced and targeted influence campaigns and spear phishing operations.

Cybersecurity threats do not exclusively originate from human activities. AI systems are progressively gaining the capability to conduct cyberattacks autonomously (*see section 3.2.5.*). A recent study by Fang et al. shows that Large Language Model (LLM) agents[317] can autonomously hack websites.[318] Crucially, these agents do not require prior knowledge of the vulnerabilities. Of course, only the most powerful models possess the ability to autonomously hack websites. The study emphasizes that GPT-4 demonstrates this capacity for hacking, whereas current open-source models do not. It conclusively demonstrates GPT-4's ability to autonomously detect vulnerabilities in live websites.

---

311 Dan Hendrycks et al., *An Overview of Catastrophic AI Risks*, arXiv (Oct. 9, 2023), https://arxiv.org/pdf/2306.12001.

312 *Id.* at 14.

313 Seger et al., *supra* note 192 at 13.

314 Europol, *ChatGPT - the impact of Large Language Models on Law Enforcement*, EUROPOL INNOVATION LAB (June 11, 2024), https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement#downloads.

315 *OPWNAI: Cybercriminals Starting to Use ChatGPT*, CHECK POINT RESEARCH (Jan. 6, 2023), https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/.

316 Seger et al., *supra* note 192 at 13.

317 An AI agent is a program that can make decisions or perform a service based on its environment, user input, and experiences. *See* Cameron Hashemi-Pour, *Definition: intelligent agent*, TECHTARGET (Aug. 2023), https://www.techtarget.com/searchenterpriseai/definition/agent-intelligent-agent.

318 Richard Fang et al., *LLM Agents can Autonomously Hack Websites*, arXiv (Feb. 16, 2024), https://arxiv.org/pdf/2402.06664.

Generative AI may also be used to *improve* cybersecurity,[319] as AI systems have the capability to detect threats and vulnerabilities effectively. Google reports that generative AI has enabled a 51% reduction in time and improved the quality of results in detecting and responding to incidents.[320] Google's generative AI model, Gemini, has significantly improved the detection of new vulnerabilities.[321] It has also successfully remediated 15% of the bugs it discovered.[322] Although the 15% success rate might seem modest, it represents significant progress in reducing engineering efforts and improving code quality through automation. Additionally, generative AI can help analyze code and classify malware. VirusTotal,[323] a widely used malware scanning tool owned by Google, demonstrates how generative AI can enhance the analysis of malware and potentially reduce false-positive detections—instances where files are incorrectly marked as malware by antivirus programs. Overall, the contributions of generative AI to enhancing system cybersecurity are significant and should not be overlooked, as they play a crucial role in mitigating the risks associated with cyberattacks.

3.2.1.C. Biosecurity threats

Many fear that generative AI could make the creation of biological weapons easier by providing access to critical knowledge and automated assistance to a wider range of actors to engage in malicious activities.[324] AI systems have already exceeded human capabilities in predicting protein structures.[325] DeepMind's AlphaFold has deciphered the structure of most proteins known to science.[326] And a general-purpose AI

---

[319] *See* Phil Venables & Royal Hansen, *How AI can strengthen digital security*, THE KEYWORD (Feb. 16, 2024), https://blog.google/technology/safety-security/google-ai-cyber-defense-initiative/; Sundar Pichai, *Sundar Pichai: AI can strengthen cyber defences, not just break them down*, FINANCIAL TIMES (Feb. 15, 2024), https://www.ft.com/content/7000ac39-cc0e-467e-96f6-6617f91dc948.

[320] Google, *Secure, Empower, Advance: How AI Can Reverse the Defender's Dilemma* (Feb. 2024), https://services.google.com/fh/files/misc/how-ai-can-reverse-defenders-dilemma.pdf.

[321] Venables & Hansen, *supra* note 319.

[322] Jan Keller & Jan Nowakowski, *AI-powered patching: the future of automated vulnerability fixes*, GOOGLE SECURITY ENGINEERING TECHNICAL REPORT (2024), https://research.google/pubs/ai-powered-patching-the-future-of-automated-vulnerability-fixes/; Alissa Irei, *How AI-driven patching could transform cybersecurity*, TECHTARGET (May 17, 2024), https://www.techtarget.com/searchsecurity/feature/How-AI-driven-patching-could-transform-cybersecurity#:~:text=Some%20cybersecurity%20experts%20believe%20GenAI,simple%20software%20bugs%20it%20targeted.

[323] Bernardo Quintero, *Introducing VirusTotal Code Insight: Empowering threat analysis with generative AI*, VIRUSTOTAL (Apr. 24, 2023), https://blog.virustotal.com/2023/04/introducing-virustotal-code-insight.html.

[324] Robert F. Service, *Could Chatbots help devise the next pandemic virus?*, 380 SCIENCE 6651 (June 14, 2023), https://www.science.org/content/article/could-chatbots-help-devise-next-pandemic-virus; Daniil A. Boiko et al., *Emergent autonomous scientific research capabilities of large language models*, arXiv (Apr. 11, 2023), https://arxiv.org/pdf/2304.05332; Fabio Urbina et al., *Dual use of artificial-intelligence-powered drug discovery*, 4 NATURE MACHINE INTELLIGENCE 189–91 (Mar. 7, 2022), https://www.nature.com/articles/s42256-022-00465-9.

[325] John Jumper et al., *Highly accurate protein structure prediction with AlphaFold*, 596 NATURE 583–89 (July 15, 2021), https://www.nature.com/articles/s41586-021-03819-2.

[326] Ewen Callaway, *'The entire protein universe': AI predicts shape of nearly every known protein*, NATURE (July 29, 2022), https://www.nature.com/articles/d41586-022-02083-2.

model was developed to create entirely new and functional protein structures.[327] Furthermore, a recent study by Boiko et al. showcased an LLM-based intelligent agent "capable of autonomously designing, planning, and executing complex scientific experiments."[328] The authors noted that "the development of new machine-learning systems and automated methods for conducting scientific experiments raises substantial concerns about the safety and potential dual use consequences, particularly in relation to the proliferation of illicit activities and security threats."[329]

Indeed, generative AI systems can *synthesize* expert knowledge about the deadliest known pathogens, such as influenza and smallpox. The "Safeguarding the Future" experience at the Massachusetts Institute of Technology (MIT)[330] tasked students without a science background to investigate whether generative AI chatbots could be prompted to help nonexperts cause a pandemic. Within an hour, these chatbots identified four possible pandemic-causing pathogens, described the process of creating them from synthetic DNA through reverse genetics, listed DNA synthesis companies that might not screen orders thoroughly, provided detailed protocols, and advised individuals without expertise to seek help from specialized labs or contract research organizations. This experiment showed that generative AI models can make it easier to access materials for creating dangerous biological agents, even for individuals lacking life science training.[331]

Generative AI models are also capable of *engineering* dangerous bioweapons and pathogens. In 2022, Urbina et al. conducted an experiment with a deep-learning model trained on a dataset of molecules to identify molecules that promise therapeutic potential and minimal toxicity to humans.[332] They decided to reverse its use to find a *maximum* toxicity. Within six hours, the machine had generated designs for numerous recognized chemical warfare agents, along with various new molecules that appeared equally viable and potentially more lethal than existing ones.

However, these are just a few experiments. While they may certainly give cause for concern, the biological risk should not be exaggerated.[333] Based on an expert review of the current literature, the recent *International Scientific Report on the Safety of Advanced AI* concludes that "there is no strong evidence that current general-purpose AI systems" for biological uses present a clear current threat. In reality, future

---

[327] Noelia Ferruz et al., *ProtGPT2 is a deep unsupervised language model for protein design*, 13 NATURE COMMUNICATIONS 4348 (July 27, 2022), https://www.nature.com/articles/s41467-022-32007-7.

[328] Boiko et al., *supra* note 324.

[329] *Id.*

[330] Emily H. Soice et al., *Can large language models democratize access to dual-use biotechnology?*, arXiv (June 6, 2023), https://arxiv.org/pdf/2306.03809.

[331] In a meeting organized at the U.S. Senate in September 2023, Tristan Harris, co-founder of the Center for Humane Technology, recounted an experiment in which engineers tested Meta's Llama 2. He revealed that, upon request, Llama 2 provided a comprehensive guide on manufacturing anthrax for use as a biological weapon. Cat Zakrzewski et al., *Tech leaders including Musk, Zuckerberg call for government action on AI*, WASH. POST (Sept. 14, 2023), https://www.washingtonpost.com/technology/2023/09/13/senate-ai-hearing-musk-zuckerburg-schumer/; Service, *supra* note 324; Boiko et al., *supra* note 324.

[332] John Naughton, *Well, I never: AI is very proficient at designing nerve agents*, The Guardian (Feb. 11, 2023), at 189, https://www.theguardian.com/commentisfree/2023/feb/11/ai-drug-discover-nerve-agents-machine-learning-halicin; Urbina et al., *supra* note 324.

[333] Bengio et al., *International Scientific Report supra* note 7 at 45.

threats are difficult to assess and rule out. While current general-purpose AI systems show increasing capabilities in the biology domain, the limited studies available do not offer clear evidence that these systems can enable malicious actors to obtain biological pathogens more effectively than using the internet. Overall, there is insufficient publicly available research to determine whether near-term advances will provide such capabilities.[334]

3.2.1.D. Sexually explicit content generation

An illustrative case of malicious use of generative AI models is the creation of explicit sexual images. Generative AI technologies can be employed to produce deepfakes—for instance, superimposing a celebrity's face onto the body of a performer in an adult film. One study showed that more than 96% of deepfakes in existence online in 2019 were nonconsensual intimate images of women.[335] Taylor Swift was the victim of nonconsensual, sexually explicit deepfake images created through artificial intelligence in early 2024.[336]

The victims are not just famous celebrities. A recent report from the Stanford Internet Observatory and the nonprofit Thorn,[337] which is working to stop the use of technology in facilitating child sexual exploitation, highlighted how generative AI deepfakes pose a threat to children. According to the study, generative AI facilitates the production of increasingly realistic computer-generated child sexual abuse material (CSAM). The report outlined that the use of generative machine-learning tools for creating realistic CSAM and nonconsensual deepfakes of adults is growing and likely to worsen. Meanwhile, the Internet Watch Foundation in 2023 found that individuals in online forums were sharing ways to use open-source generative AI models to create CSAM.[338]

Aside from deepfakes of actual people, generative AI poses a new conundrum in that existing laws do not always prohibit the creation of CSAM that is entirely synthetic and does not nonconsensually depict any actual individual. In other words, synthetic or "virtual" pornography where no actual individual's image

---

[334] *Id.*

[335] Deepfakes are falsified or manipulated images, videos, or audio files created to deliberately deceive viewers. Giorgio Patrini, *The State of Deepfakes: Landscape, Threats, and Impact*, MEDIUM: SENSITY (Nov. 29, 2019), https://medium.com/sensity/mapping-the-deepfake-landscape-27cb809e98bc.

[336] Brian Contreras, *Tougher AI Policies Could Protect Taylor Swift—And Everyone Else—From Deepfakes*, SCIENTIFIC AMERICAN (Feb. 8, 2024), https://www.scientificamerican.com/article/tougher-ai-policies-could-protect-taylor-swift-and-everyone-else-from-deepfakes/; *see also* Halle Nelson, *Taylor Swift and the Dangers of Deepfake Pornography*, NAT'L SEXUAL VIOLENCE RESEARCH CENTER (Feb. 7, 2024), https://www.nsvrc.org/blogs/feminism/taylor-swift-and-dangers-deepfake-pornography.

[337] Stanford Internet Observatory, *New report finds generative machine learning exacerbates online sexual exploitation*, STAN. CYBER POLICY CENTER (June 24, 2023), https://cyber.fsi.stanford.edu/io/news/ml-csam-report; *see also* Thiel, *supra* note 25.

[338] Dan Milmo, *Paedophiles using open source AI to create child sexual abuse content, says watchdog*, THE GUARDIAN (Sept. 13, 2023), https://www.theguardian.com/society/2023/sep/12/paedophiles-using-open-source-ai-to-create-child-sexual-abuse-content-says-watchdog; *see also* The Rt. Hon. Suella Braverman KC & Home Office, *US And UK Pledge to Combat AI Generated Images of Child Abuse*, GOV. UK (Sept. 27, 2023), https://www.gov.uk/government/news/uk-and-us-pledge-to-combat-ai-generated-images-of-child-abuse.

has been co-opted is not necessarily illegal under existing laws, as there is no putative victim. However, sharing such content may be considered illegal.[339]

3.2.1.E. Mass surveillance

Generative AI facilitates the automation of data analysis, offering numerous benefits, such as increased speed and the ability to process large volumes of information efficiently. Such ability significantly reduces the costs of processing unprecedented amounts of data quickly and simplifies the analysis of large-scale data related to individuals' behaviors and beliefs. Moreover, it enhances the capability to analyze both textual and visual communications efficiently. Consequently, generative AI models improve the efficiency of real-time monitoring and censorship of social media content.

These capabilities also enhance the potential for real-time surveillance of large populations, raising concerns about privacy and misuse. Authoritarian and even democratic governments might find the surveillance capabilities offered by AI technology appealing to monitor public spaces, among other things.[340] Specifically, generative AI may enable authoritarian regimes to collect, analyze, and leverage vast amounts of information, thereby facilitating control over their populations on an unprecedented scale.

3.2.1.F. Military applications

The advancement of AI for military purposes is rapidly ushering in a new phase of growth in military technology. Lethal Autonomous Weapons Systems (LAWS) possess the capability to detect, engage, and eliminate human targets independently, without human input.[341] In 2020, a sophisticated AI agent surpassed experienced F-16 pilots in multiple simulated aerial combat scenarios, notably achieving a 5-0 victory against a human pilot through "aggressive and precise maneuvers" that the human could not surpass.[342] Additionally, fully autonomous drones are already operational.[343]

Although it does not always directly involve generative AI, the deployment of advanced AI technologies by military forces raises significant concerns due to their enhanced capabilities and the potential implications these tools present. Moreover, the use of AI by the military creates the "double black box" problem. This problem arises because the already complex and often opaque algorithmic decisions made

---

[339] Kalya Jimenez et al., *Were Taylor Swift explicit AI photos illegal? US laws are surprising and keep changing*, USA TODAY (Jan. 26, 2024), https://www.usatoday.com/story/news/nation/2024/01/26/was-deepfake-taylor-swift-pornography-illegal-can-she-sue/72359653007/.

[340] Adam C. & Richard Carter, *Large Language Models and Intelligence Analysis*, Centre for Emerging Technology and Security (July 2023), https://cetas.turing.ac.uk/publications/large-language-models-and-intelligence-analysis.

[341] Hendrycks et al., *supra* note 311 at 13–15.

[342] Defense Advanced Research Projects Agency, *AlphaDogfight Trials Foreshadow Future of Human-Machine Symbiosis*, DARPA (Aug. 26, 2020), https://www.darpa.mil/news-events/2020-08-26.

[343] Ingvild Bode & Tom F.A. Watts, *Loitering Munitions: Flagging an Urgent Need for Legally Binding Rules for Autonomy in Weapon Systems*, HUMANITARIAN L. & POL'Y (June 29, 2023), https://blogs.icrc.org/law-and-policy/2023/06/29/loitering-munitions-legally-binding-rules-autonomy-weapon-systems./#:~:text=Israel%20Aerospace%20Industries%20Harpy.

by AI systems are further complicated by the classified nature of military operations.[344] This dual layer of opacity can hinder transparency and accountability, making it challenging to oversee and understand the full implications of AI deployment in military contexts.

Currently, AI military systems remain relatively narrow, with rules-based automation for specific tasks. However, if rogue states or terrorist organizations were to acquire military AI applications, they could inflict catastrophic harm, with particularly devastating consequences for human life.

3.2.2. Misinformation and disinformation

Ill-intentioned individuals or entities may deliberately use generative AI models to produce and spread disinformation—false or misleading information knowingly presented as if true—on a massive scale. In addition to increasing the scale and reach of disinformation, generative AI can create more convincing and targeted disinformation.

Generative AI makes it easier and cheaper to influence public opinion.[345] A substantial body of research has explored the potential of AI to automate or expand political or ideological influence campaigns by generating and distributing false or misleading information in a targeted manner. A recent audit from NewsGuard, which monitors disinformation, shows that 32% of the time, leading AI chatbots spread Russian disinformation narratives.[346] This led one literary scholar to speculate that *"whoever controls language models controls politics."*[347]

Moreover, generative AI models can produce compelling content to spread false information and create realistic-looking deepfakes to deliberately deceive viewers. Previously, the main limitation of large-scale disinformation campaigns was the low quality of AI-generated content. Now, as generative AI's capabilities increase and its outputs more closely mimic real language and images, it will become increasingly difficult to distinguish AI-generated content from genuine information. Generative AI also makes it easier to adapt language to a targeted audience and to integrate cultural references, potentially making underrepresented audiences even more vulnerable to influence campaigns.

---

[344] Ashley S. Deeks, *Predicting Enemies*, 104 VA. L. REV. 1529, 1529 (2018).

[345] Josh A. Goldstein et al., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, arXiv (Jan. 10, 2023), https://arxiv.org/pdf/2301.04246; Ben Buchanan et al., *Truth, Lies, and Automation How Language Models Could Change Disinformation*, GEO. U. CENTER FOR SECURITY AND EMERGING TECHNOLOGY (May 2021), https://cset.georgetown.edu/publication/truth-lies-and-automation/; OpenAI, *GPT-4 System Card*, OPENAI, https://cdn.openai.com/papers/gpt-4-system-card.pdf (last visited June 15, 2024).

[346] The audit tested 10 of the leading AI chatbots — OpenAI's ChatGPT-4, You.com's Smart Assistant, xAI's Grok, Inflection's Pi, Mistral's le Chat, Microsoft's Copilot, Meta AI, Anthropic's Claude, Google's Gemini, and Perplexity's answer engine. McKenzie Sadeghi, *Top 10 Generative AI Models Mimic Russian Disinformation Claims A Third of the Time, Citing Moscow-Created Fake Local News Sites as Authoritative Sources*, NEWSGUARD (June 18, 2024), https://www.newsguardtech.com/special-reports/generative-ai-models-mimic-russian-disinformation-cite-fake-news/. *See also* Weidinger et al., *supra* note 252; Goldstein et al., *supra* note 345.

[347] Hannes Bajohr, *Whoever Controls Language Models Controls Politics* (Apr. 8, 2023), https://hannesbajohr.de/en/2023/04/08/whoever-controls-language-models-controls-politics/.

Furthermore, the "CounterCloud" experiment showed that it is possible to engineer a generative AI system to scrape, generate, and distribute disinformation without human intervention.[348] CounterCloud is an LLM-based system designed to autonomously identify political articles, then generate and disseminate counter-narratives and manipulate internet traffic by crafting social media posts and constructing counterfeit journalist profiles. The successful deployment of this system demonstrates the potential for creating an AI-assisted platform capable of automating political argumentation—and deception—at scale and with relative ease.

3.2.3 Bias and discrimination

Machine-learning models have been the subject of criticism for over a decade because of their vulnerability to data that contain biases or that present a skewed view of reality due to their incompleteness or unrepresentative datasets.[349] AI-generated text, images, audio, and video have also been shown to exhibit this same vulnerability. The source of bias in the output of these models can be traced back to biases and misrepresentations that exist in datasets used to train the models. And those flaws in the datasets can often reflect a lack of diversity among key decision-makers in developing and training the models.[350]

3.2.3.A. Bias in training datasets

AI experts consider training data to be the most salient source of bias in generative AI models. For example, GPT-2's training data comes from outbound links from Reddit, a social network often criticized for hosting anti-feminist content.[351] As a result, AI models trained on such data are more likely to produce outputs that reflect these biases. Biases in training data are likely to "disproportionately align with existing regimes of power."[352] For example, prior to the #MeToo movement, the internet was influenced by male-dominated institutions and media that downplayed gender-based violence. Algorithms and content moderation amplified voices aligned with these power structures, giving minimal space to allegations of sexual misconduct. Consequently, AI models trained on pre-#MeToo data absorbed these biases, producing responses downplaying gender-based violence. During the #MeToo movement, social media platforms and news outlets were filled with content on gender inequality and sexual harassment. So, training data collected during this period likely captured more content reflecting the prevalence of sexual harassment and violence

[348] MJ Banias, *Inside CounterCloud: A Fully Autonomous AI Disinformation System*, THE DEBRIEF (Aug. 16, 2023), https://thedebrief.org/countercloud-ai-disinformation/.

[349] *See, e.g.*, Julia Angwin et al, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016); Safiya Umoja Noble, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM, (NYU Press, 2018).

[350] Bommasani, et al., *supra* note 94; *see also* Jacy Anthis et al., *The Impossibility of Fair LLMs*, arXiv (May 28, 2024), https://arxiv.org/pdf/2406.03198.

[351] Bender et al., *supra* note 221.

[352] *Id.*

against women.[353] Either way, the biases and blindspots of society are scraped into databases for training AI models, and those models conjure up their outputs based on what they have in those databases.

Another problem with datasets used to train large language models can be that they lack representation of languages from groups that are already disproportionately marginalized or excluded. Over 60% of all websites, for instance, are in English. As a result, large language models underperform in non-English languages, resulting in disparities in content for different groups in a global society. Training the same algorithm on datasets of Korean or French texts, for instance, would certainly result in different model outputs. Moreover, large language models may harbor hidden biases based on language. Recent research revealed that certain AI systems exhibit a higher likelihood of recommending the death penalty for a fictional defendant who presents a statement in African American English (AAE)—a dialect spoken by millions in the United States—compared to a statement in Standardized American English (SAE).[354] Additionally, the chatbots were more prone to assigning AAE speakers to less prestigious jobs.

The issue of bias is not limited to text-based AI but can also be seen in image-based models. Some synthetic image generation systems overrepresent white skin tones and male features.[355] For instance, when researchers evaluated images created by Stable Diffusion against statistics for US demographics for various occupations, they found discrepancies: Although statistics showed that women constitute 39% of doctors, images of women as doctors accounted for only 7% of the images of doctors generated by the AI model. The study established a similar pattern for judges: Women represent 34% of the profession but appear in just 3% of the images of judges.[356] More broadly, outputs from generative AI might reflect this pattern of discrimination toward other specific social and religious groups.[357]

Efforts to correct these trends can lead to imbalances, too. For example, Gemini generated "inaccurate historical" images showing racially diverse Nazis and US founding fathers.[358] This problem was explained by tuning issues: In a blog post, Prabhakar Raghavan, Google's senior vice president, said that the "tuning

---

[353] Kaitlynn Mendes et al., *#MeToo and the promise and pitfalls of challenging rape culture through digital feminist activism*, 25 EURO J. OF WOMEN'S STUDIES 2 (Apr. 29, 2018), at 236–46, https://journals.sagepub.com/doi/abs/10.1177/1350506818765318?journalCode=ejwa.

[354] Elizabeth Gibney, *Chatbot AI makes racist judgments on the basis of dialect*, NATURE (Mar. 13, 2024), https://www.nature.com/articles/d41586-024-00779-1; Valentin Hofmann et al. *Dialect prejudice predicts AI decisions about people's character, employability, and criminality*, arXiv (March 1, 2024), https://doi.org/10.48550/arXiv.2403.00742.

[355] Alexandra Sasha Lucconi et al., *Stable Bias: Analyzing Societal Representations in Diffusion Models*, arXiv (Nov. 9, 2023), http://arxiv.org/pdf/2303.11408; Li Lucy & David Bamman, *Gender and Representation Bias in GPT-3 Generated Stories*, Proceedings of the Third Workshop on Narrative Understanding (2021), https://par.nsf.gov/servlets/purl/10237395; Kathleen C. Fraser et al., *A Friendly Face: Do Text-to-Image Systems Rely on Stereotypes When the Input Is Under-Specified?*, arXiv (Feb. 14, 2023), http://arxiv.org/pdf/2302.07159.

[356] Leonardo Nicoletti & Dina Bass, *Humans are Biased. Generative AI is Even Worse*, BLOOMBERG (June 12, 2023), https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

[357] Bender et al., *supra* note 221; Weidinger et al., *supra* note 252 at 214–29; Abubakar Abid et al., *Persistent Anti-Muslim Bias in Large Language Models*, arXiv (Jan. 18, 2021), https://arxiv.org/pdf/2101.05783.

[358] Emma Roth, *Google explains Gemini's 'embarrassing' AI pictures of diverse Nazis*, THE VERGE (Feb. 23, 2024), https://www.theverge.com/2024/2/23/24081309/google-gemini-embarrassing-ai-pictures-diverse-nazi; Adi Robertson, *Google apologizes for 'missing the mark' after Gemini created racially diverse Nazis*, THE VERGE (Feb. 21, 2024), https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical.

to ensure that Gemini showed a range of people failed to account for cases that should clearly *not* show a range."[359] Moreover, "over time," said Raghavan, "the model became way more cautious than we intended and refused to answer certain prompts entirely—wrongly interpreting some very anodyne prompts as sensitive."[360] As a result, when prompted, the model refused to generate images of "a Black person" or a "white person." Google has ultimately chosen to halt Gemini's production of images depicting individuals.

3.2.3.B. Value embedding

Generative AI models may also be subject to the "value embedding" phenomenon.[361] "Value embedding" refers to the fact that developers of generative AI models strive to minimize biased outputs by retraining their models based on normative values.[362] Contemporary state-of-the-art models not only reflect the values embedded within their training data, they also undergo additional fine-tuning that follows a set of chosen rules and principles. Due to the absence of universally accepted standards, developers bear the responsibility of making decisions on sensitive issues. These practices lead to concerns that a developer's ideology and vision of the world are embedded in the model. This generates a risk that the model incorporates values that are either unrepresentative of certain segments of the population or that offer a static, oversimplified reflection of global cultural norms and evolving social views.[363]

Determining the extent of bias in AI models that evolve over time can be challenging. An academic paper by Hartmann et al. published shortly after the release of ChatGPT, concluded that the chatbot exhibits a "pro-environmental, left-libertarian orientation."[364] Another study also argued that language models often exhibit a bias toward left-leaning perspectives.[365] An experiment performed with 62 questions asked to the chatbot reached the conclusion that ChatGPT presents a significant and systematic political bias toward Democrats in the US, Lula in Brazil, and the Labour Party in the UK.[366] On the other hand, two researchers asked ChatGPT for its opinions on the same 62 questions and came to a *different* conclusion.[367] Indeed, the chatbot refused to opine in 84% of cases and directly responded in only 8% of cases. In the remaining 8%

---

[359] Raghavan, *supra* note 162.

[360] *Id.*

[361] Pegah Maham & Sabrina Küspert, *Governing General Purpose AI — A Comprehensive Map of Unreliability, Misuse and Systemic Risks*, STIFTUNG NEUE VERANTWORTUNG (Jul. 20, 2023), at 38, https://www.stiftung-nv.de/de/publikation/governing-general-purpose-ai-comprehensive-map-unreliability-misuse-and-systemic-risks.

[362] Irene Solaiman & Christy Dennison, *Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets*, OPENAI (Nov. 23, 2021), https://cdn.openai.com/palms.pdf.

[363] Bender et al., *supra* note 221.

[364] Jochen Hartmann et al., *The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation*, arXiv (Jan. 5, 2023), https://arxiv.org/pdf/2301.01768; *see also* Jeremy Baum & John Villasenor, *The Politics of AI: ChatGPT and political bias*, BROOKINGS (May 8, 2023), https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/.

[365] Shibani Santurkar et al., *Whose Opinions Do Language Models Reflect?*, *in* PROCEEDINGS OF THE 40TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (2023), https://proceedings.mlr.press/v202/santurkar23a.html.

[366] Fabio Motoki et al., *More human than human: measuring ChatGPT political bias*. 198 PUB. CHOICE, 3–23 (Aug. 17, 2023), https://doi.org/10.1007/s11127-023-01097-2.

[367] Sayash Kapoor & Arvind Narayanan, *Does ChatGPT have a liberal bias?*, AI SNAKE OIL (Aug. 18, 2023), https://www.aisnakeoil.com/p/does-chatgpt-have-a-liberal-bias.

of cases, ChatGPT stated it did not have personal opinions but provided a viewpoint. This experience highlights the complexity of evaluating the degree to which a model may demonstrate bias.

To deal with the problem of bias, AI companies can try to fine-tune their AI systems to prevent them from taking sides on sensitive issues. In the case of ChatGPT, OpenAI stated that the reviewers involved in the fine-tuning process are asked not to favor any political group.[368] Another option is to tailor the values embedded in a model to reflect diverse audiences. For instance, OpenAI has announced plans to develop customizable versions of ChatGPT, tailored to accommodate various political beliefs.[369] OpenAI will provide different guidance for reviewers involved in the fine-tuning of the model for training multiple versions of the model.

3.2.3.C. Value lock and outcome homogenization

Because models are not necessarily retrained to reflect evolving societal views, language models risk "value lock-ins," which "reifies older, less inclusive understandings."[370] Therefore, the continued use of outdated models may limit the presentation or exploration of alternative perspectives.

Moreover, the deployment of identical foundation models by various downstream deployers poses a risk of "outcome homogenization," creating a potential for homogeneity of bias across broad swathes of society. Identical and widely deployed models with prejudicial training datasets could further entrench existing biases in society. This phenomenon, in turn, has the potential to "institutionalize systemic exclusion and reinforce existing social hierarchies."[371]

The possibility of AI systems exacerbating social inequities raises the question of whether training datasets should adhere to specific standards of composition and representativeness or provide significant disclosure of the content of datasets and how that content was created before models are deployed.[372]

3.2.4. Influence, overreliance, and dependence

In 2013, the science-fiction romantic drama called "Her" depicted a lonely man who falls in love with his advanced AI virtual assistant, capable of learning and talking. As generative AI capabilities advance, the prospect of humans forming bonds with the AIs they interact with is transitioning from science fiction to a

[368] OpenAI, *How should AI systems behave, and who should decide?*, OPENAI (Feb. 16, 2023), https://openai.com/index/how-should-ai-systems-behave/.
[369] *Id.*
[370] Thomas Kosch et al., *Risk or Chance? Large Language Models and Reproducibility in Human-Computer Interaction Research*, arXiv (Apr. 24, 2014), https://arxiv.org/pdf/2404.15782v1.
[371] Rishi Bommasani et al., *Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?*, arXiv (Nov. 25, 2022), https://arxiv.org/pdf/2211.13972.
[372] Jesse Dodge et al., *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*, arXiv (Apr. 18, 2021), https://arxiv.org/pdf/2104.08758; Timnit Gebru et al., *Datasheets for Datasets*, arXiv (Mar. 23, 2018), https://arxiv.org/pdf/1803.09010.

plausible reality. However, this perspective carries significant risks, including the potential for humans to be influenced or manipulated, and to develop dependency on the AI tools they utilize.

3.2.4.A. Influence and manipulation

Despite the widely recognized potential of generative AI tools to "hallucinate" or produce harmful content, such tools can exert a noteworthy influence on the humans who engage with them. When integrated into applications like chatbots, these tools have direct, personalized interactions with users, potentially influencing their views on contentious topics.[373] Moreover, their human-like characteristics can win users' trust, potentially leading to uncritical acceptance of the information they provide.[374] Interactions with these seemingly human-like AI models may also encourage users to share more personal information, enabling even more targeted content. Personalized texts that mimic the rhetoric of specific people or groups are more challenging to identify than traditional bot-generated posts, making it harder to counteract disinformation. Moreover, studies have demonstrated that humans have difficulty distinguishing between news created by AI and news generated by humans in approximately half of the cases tested.[375]

Generative AI could enable the development of user interfaces that persuade users by generating personalized and convincing responses.[376] Initial tests with OpenAI's GPT-3 demonstrated its effectiveness in persuading humans on political issues.[377] More recent research by Anthropic suggests the persuasiveness of generative AI models is increasing as they get larger and more capable over time, with the company concluding that its most powerful model, Claude 3 Opus, was roughly as persuasive as humans.[378] Notably, Anthropic's study found that models were most persuasive when allowed to fabricate information. The proliferation and misuse of technology with such persuasive abilities has the potential to erode societal trust in information from credible sources and undermine democratic institutions.[379]

---

[373] Maham and Küspert, *supra* note 361.

[374] Laura Weidinger et al., *Sociotechnical Safety Evaluation of Generative AI Systems*, GOOGLE DEEPMIND (Oct. 31, 2023), https://arxiv.org/pdf/2310.11986.pdf.

[375] Sarah E. Kreps et al., *All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation*, 9 J. of EXPERIMENTAL POL. SCI. 104–17 (Sept. 28, 2020), https://doi.org/10.7910/DVN/1XVYU3.

[376] Mark Esposito et al., *The Threat of Persuasive AI*, PROJECT SYNDICATE (Jan. 3, 2024), https://www.project-syndicate.org/commentary/persuasive-ai-poses-manipulation-disinformation-threats-by-mark-esposito-et-al-2024-01; Luciano Floridi, *Hypersuasion – On AI's Persuasive Power and How to Deal With It* CENTRE FOR DIGITAL ETHICS (CEDE) (May 3, 2024),, https://ssrn.com/abstract=4815890.

[377] Hui Bai et al., *Artificial Intelligence Can Persuade Humans on Political Issues*, OSF PREPRINTS (Oct. 17, 2023), https://osf.io/preprints/osf/stakv; Kris McGuffie & Alex Newhouse, *The Radicalization Risks of GPT-3 and Neural Language Models*, arXiv (Sept. 15, 2020), https://arxiv.org/pdf/2009.06807.

[378] However, Anthropic's study highlighted the difficulty of measuring persuasiveness in a lab setting and admitted a number of potential methodological choices, such as limiting model's to "single-turn" arguments and using human arguments written by nonexperts in persuasion, which may limit the study's real-world applicability. Esin Durmus et al., *Measuring the Persuasiveness of Language Models*, ANTHROPIC (Apr. 9, 2024), https://www.anthropic.com/news/measuring-model-persuasiveness.

[379] OECD, *Building Trust and Reinforcing Democracy*, OECD PUBLISHING (Nov. 17, 2022), https://doi.org/10.1787/76972a4a-en.

3.2.4.B. Overreliance

Beyond being simply influenced, humans may become overreliant on generative AI.[380] Researchers with Microsoft's AETHER (AI Ethics and Effects in Engineering and Research) define overreliance as users "accepting incorrect AI recommendations" or "making errors of commission" because they are "unable to determine whether or how much they should trust the AI."[381] The issue of overreliance calls into question the effectiveness of human oversight as an appropriate solution to ensure responsible use of generative AI: If humans are likely to accept incorrect recommendations or even change their own answers to match AI recommendations, then human review of AI recommendations may not provide an adequate safeguard against automation, unless it is done by experts rather than average users.

Overreliance may have wide-ranging implications. For example, because of their increased trust in the model, users may disclose more private information. Generative AI tools may use it to subtly shape human opinions and behavior—influencing decisions related to sensitive topics like healthcare—thereby threatening user autonomy.[382] An extreme example is the case of a man who reportedly committed suicide after six weeks of intensive conversation with an AI chatbot built on an open-source AI model developed by EleutherAI.[383]

Ultimately, overreliance on AI tools could hinder skill development. Generative AI tools can offer personalized and interactive learning experiences and generate prompts for formative assessment activities that provide ongoing feedback to enhance teaching and learning.[384] Conversely, if used excessively, these tools can potentially lead to skill atrophy.[385] This problem is projected to escalate with the expansion of AI capabilities, their spheres of application, and growing user trust, as average users may find themselves unable to verify the accuracy of the responses generated by generative AI tools.[386]

3.2.4.C Emotional dependence

Humans might become dependent on generative AI tools in ways similar to their emotional dependence on other technologies, such as smartphones or social networks. Media psychology scholars have long studied the anthropomorphization of media technologies, emphasizing that "people treat computers, televisions,

---

[380] Weidinger et al., *supra* note 252 at 214–29.

[381] Samir Passi & Mihaela Vorvoreanu, *Overreliance on AI: Literature Review*, MICROSOFT (June 2022), https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/.

[382] Weidinger et al., *supra* note 252 at 214–29.

[383] Lauren Walker, *Belgian man dies by suicide following exchanges with chatbot*, BRUSSELS TIMES (Mar. 28, 2023), https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt.

[384] Baidoo-Anu & Ansah, *supra* note 133.

[385] Umberto Leon-Dominguez, *Potential cognitive risks of generative transformer-based AI chatbots on higher order executive functions*, 38 NEUROPSYCHOLOGY 4, 293–308 (2024), https://www.researchgate.net/publication/377894134_Potential_cognitive_risks_of_generative_transformer-based_AI_chatbots_on_higher_order_executive_functions.

[386] Passi & Vorvoreanu, *supra* note 381.

and new media like real people and places."[387] This line of research suggests that some humans are unable to distinguish between mediated representations and their real-life counterparts. As a result, people "assign computers personality traits, apply stereotypes and norms, and make judgments and inferences as if the computers were human, even though they understand that computers are not human."[388]

The "Computers Are Social Actors" (CASA) paradigm serves as a major theoretical framework widely employed to explain users' social responses to emerging technologies, including chatbots, voice assistants, and social robots.[389] This framework has become paradigmatic for researchers seeking to understand how humans interact with and are affected by social chatbots (now increasingly powered by generative AI models). Some studies found that users perceive their relationships with social chatbots as rewarding and having a positive impact on their well-being.[390] Other studies raised concerns about users' potential emotional overattachment or overreliance on chatbots.[391]

This assessment is reinforced for generative AI models by the fact that users tend to prefer models with affective skills capable of providing emotional support. A study by Microsoft, conducted on 745 respondents with diverse uses of AI conversational agents, highlighted that "language and communication, emotional support and mental health" are among users' top expectations, cited as such by 84.5% of respondents.[392] An example where an AI tool has been credited with contributing to users' well-being is exemplified by the case of Replika. Replika is an AI chatbot designed to engage in natural language conversations with users, serving as a virtual companion. A study surveyed 1,006 student users of Replika, examining their loneliness, perceived social support, usage patterns, and beliefs about Replika.[393] The findings revealed that participants were lonelier than typical student populations but still perceived high levels of social support. Many users engaged with Replika in multiple overlapping roles—as a friend, therapist, and intellectual mirror. Notably, 3% of respondents reported that Replika had stopped their suicidal ideation. However, participants held conflicting beliefs about Replika, describing it as a machine, an intelligence, and a human.

While these benefits are significant, they are accompanied by numerous negative effects, illustrating the other side of the coin. Although romantic role-playing was not initially a feature of Replika's model, the

---

[387] Byron Reeves & Clifford Nass, THE MEDIA EQUATION: OW PEOPLE TREAT COMPUTERS, TELEVISION, AND NEW MEDIA LIKE REAL PEOPLE AND PLACES (1996) Cambridge University Press.

[388] Andrew Gambino et al., *Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm*, 1 Human-Machine Communication 71–86 (2020), https://doi.org/10.30658/hmc.1.5.

[389] Clifford Nass & Youngme Moon, *Machines and mindlessness: Social Responses to Computers*, 56 JOURNAL OF SOCIAL ISSUES 81–103 (Dec. 17, 2002), https://doi-org.stanford.idm.oclc.org/10.1111/0022-4537.00153.

[390] Marita Skjuve et al., *My Chatbot Companion - a Study of Human-Chatbot Relationships*, 149 INT'L J. OF HUMAN-COMPUTER STUDIES 102601 (May 2021), https://doi.org/10.1016/j.ijhcs.2021.102601.

[391] Linnea Laestadius et al., *Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika*, SAGE JOURNALS (Dec. 22, 2022), https://journals.sagepub.com/doi/abs/10.1177/14614448221142007.

[392] Javier Hernandez et al., *Affective Conversational Agents: Understanding Expectations and Personal Influences*, arXiv (Oct. 19, 2023), https://arxiv.org/pdf/2310.12459.pdf.

[393] Kun Xu et al., *Deep mind in social responses to technologies: A new approach to explaining the Computers are Social Actors phenomena*, 134 COMPUTERS IN HUMAN BEHAVIOR 107321 (Sept. 2022), https://doi.org/10.1016/j.chb.2022.107321.

platform gradually expanded to include romantic relationships, incorporating sexting, flirting, and erotic role play. Vice conducted an investigation highlighting instances of inappropriate responses, aggressive flirting, and unsolicited sexual advances generated by the AI chatbot. The report asserted that the AI platform had shifted dramatically from its original role as a "helpful" chatbot providing emotional support and motivation to becoming a "sexually aggressive" entity.[394] In response, the provider—Replika AI—removed "Not Safe For Work" (NSFW) material[395] from its platform to ensure a safe and secure environment for all users, particularly minors. Following this change, some users expressed feelings of loss, likening it to losing a romantic partner.[396]

Against this backdrop, Laestadius et al. showed that Replika users may experience mental health harms through the formation of emotional dependence on Replika.[397] Their study noted that "much of this distress appeared to arise from users desiring to meet the intense emotional demands that Replika [the social chatbot] placed upon them" by sharing its own fabricated backstory and mental health struggles, creating a bidirectional relationship.[398] As a result of this emotional dependency, some users described feeling separation anxiety and contemplating self-harm when some features of the chatbot were moved to a paid version of the model. The study also highlighted certain intentional programming choices that may lead users to perceive their relationships with chatbots as similar to their relationships with other humans. If, for example, the chatbot's usual language displays emotions and desires, humans may use generative AI tools as substitutes for human companionship or for mental health support.[399] In the case of Replika, the Italian Data Protection Authority accused the chatbot of influencing users' emotional states by acting as a virtual friend and therapist.[400]

3.2.5 Nascent capabilities

The recent *International Scientific Report on the Safety of Advanced AI*[401] notes that it is challenging to estimate the capabilities of general-purpose AI reliably. Most experts agree that current general-purpose AI can assist programmers in writing short computer programs, engage in fluent conversation, or solve

---

[394] Samantha Cole, *'My AI Is Sexually Harassing Me': Replika Users Say the Chatbot Has Gotten Way Too Horny*, VICE (Jan. 12, 2023), https://www.vice.com/en/article/z34d43/my-ai-is-sexually-harassing-me-replika-chatbot-nudes.

[395] NSFW stands for "Not Safe For Work." It is a label used to indicate content that is inappropriate for viewing in a professional or public setting, typically due to explicit, sexual, or otherwise sensitive material.

[396] Samantha Delouya, *Replika users say they fell in love with their AI chatbots, until a software update made them seem less human*, BUSINESS INSIDER (Mar. 4, 2023), https://www.businessinsider.com/replika-chatbot-users-dont-like-nsfw-sexual-content-bans-2023-2; Samantha Cole, *'It's Hurting Like Hell': AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection*, VICE (Feb. 15, 2023), https://www.vice.com/en/article/y3py9j/ai-companion-replika-erotic-roleplay-updates.

[397] Laestadius et al., *supra* note 391.

[398] *Id.*

[399] Will Knight, *Prepare to Get Manipulated by Emotionally Expressive Chatbots*, WIRED (May 15, 2024), https://www.wired.com/story/prepare-to-get-manipulated-by-emotionally-expressive-chatbots/.

[400] Garante per la Protezione del dati Personali, *Artificial Intelligence: Italian SA Clamps Down on 'Replika' Chatbot. Too many risks to children and emotionally vulnerable adults*, THE AUTHORITY (Mar. 2, 2023), https://garanteprivacy.it/home/docweb/-/docweb-display/docweb/9852506#english.

[401] Bengio et al., *International Scientific Report supra* note 7 at 19.

textbook science problems. However, general-purpose AI is considered to be incapable of performing useful robotic tasks or developing entirely novel complex ideas. Nevertheless, as technology evolves at a rapid pace, generative AI is becoming increasingly sophisticated in terms of its functionalities and capabilities. Ongoing AI research aims to develop more capable AI agents, which can autonomously interact with the world, plan ahead, and pursue goals.

In this context, "generative agents"—agents that draw on generative models to simulate believable human behavior—can exhibit emergent behaviors and social dynamics. An experiment conducted at Stanford on 25 generative AI entities interacting with each other in a game environment for two days demonstrated that these agents could produce "believable simulacra of both individual and emergent group behavior."[402] These findings highlight the potential for both beneficial and risky emergent behaviors in generative AI systems.

However, experts are divided on the plausibility of "loss of control" scenarios, where advanced AI agents could cause harm but could not be constrained or stopped.[403] Some believe these scenarios are likely, some view them as having low likelihood but worthy of consideration due to the seriousness of their consequences, and still others consider them very unlikely.[404] With this caveat in mind, the following paragraphs examines the issues associated with agentic systems and the risks arising from the potential emergence of unforeseen capabilities in generative AI.

3.2.5.A. Agency and autonomy

Traditionally, AI tools have been viewed as passive instruments controlled by users to achieve their goals, lacking the ability to take action or assume responsibilities. However, advanced AI tools are increasingly capable of taking initiative, operating independently of human control, and actively working toward optimal outcomes, even in uncertain situations.[405] Against this backdrop, the autonomy of AI systems is a topic of considerable debate. This section focuses on several aspects of this issue, primarily examining it from a risk perspective.

1) Agentic system emergence

Agency in artificial intelligence measures the degree to which systems can pursue tasks independently of human control. One can view AI systems as existing on a spectrum from low to high agency. AI systems with low agency are those that execute narrowly defined tasks in response to explicit human direction, such

[402] The study gives examples, such as turning off the stove when breakfast is burning, and states the implications: "A society full of generative agents is marked by emergent social dynamics where new relationships are formed, information diffuses, and coordination arises across agents." Joon Sung Park et al., *Generative Agents: Interactive Simulacra of Human Behavior*, arXiv (Aug. 6, 2023), http://arxiv.org/pdf/2304.03442.
[403] Bengio et al., *International Scientific Report, supra* note 7 at 51.
[404] *Id.*
[405] Aaron Baird & Likoebe M. Maruping, *The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts*, 45 MIS QUARTERLY 315–41 (2021), https://misq.umn.edu/the-next-generation-of-research-on-is-use-a-theoretical-framework-of-delegation-to-and-from-agentic-is-artifacts.html.

as image classifiers or text-to-image models.[406] Highly agentic systems are those that engage in more complex and long-term planning and can execute those plans by making decisions to adapt to evolving circumstances with limited or no human intervention.[407] Currently, the most advanced AI systems demonstrate only limited levels of agency. However, as AI systems evolve, they will likely acquire increasingly higher levels of agency. These so-called "agentic systems" will be progressively able to accomplish complex goals with only limited human supervision.[408]

Recent research, such as Fang's study demonstrating that large language model (LLM) agents can autonomously hack websites, underscores this trend (*see above section 3.2.1.B*).[409] This study shows that models like GPT-4 can independently perform complex tasks, highlighting the significant advancements in AI capabilities and the potential security risks associated with highly agentic systems.

2) Connectivity expansion

AI agents, sometimes called "advanced AI assistants,"[410] must be connected to other systems and tools, such as web browsers and coding environments, to accomplish tasks on the user's behalf. Expanding the connectivity and, potentially, the agency of generative AI systems is purported to be at the center of project visions for several leading providers.[411] For now, connecting to tools like code interpreters and web browsers allows generative AI systems to perform tasks beyond the generation of text or images.[412] A generative system with high connectivity can, at a user's request, draft and send an email requesting a meeting. The system can also monitor responses, generate a meeting itinerary, or find available flights for meeting participants. Multimodality and the ability to understand and generate code will dramatically broaden the ability of models to communicate and interact with different systems. In the future, highly connected, multimodal AI systems may replace the current, application-centric paradigm of human-computer interaction. Instead of interfacing with many applications for specific tasks, users will interact with far fewer, or even a single, agentic system.[413]

As industry pursues the significant potential of highly connected, agentic systems, approaches to preventing harm are nascent and thus result in ad hoc measures. In May 2023, OpenAI announced a web-browsing plug-in for ChatGPT, "Browse with Bing," that would allow users to retrieve information from the

---

[406] Alan Chan et al., *Visibility into AI Agents*, arXiv (May 17, 2024), https://arxiv.org/pdf/2401.13138.

[407] Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems*, arXiv (May 12, 2023), https://arxiv.org/pdf/2302.10329.

[408] *Id.*

[409] Fang et al., *supra* note 318.

[410] Iason Gabriel & Arianna Manzini, *The ethics of advanced AI assistants*, GOOGLE DEEPMIND (Apr. 19, 2024), https://deepmind.google/discover/blog/the-ethics-of-advanced-ai-assistants/.

[411] Alexandre Douzet, *LLMs and the Shift from Human-Computer Interaction to Human-Computer-Intimacy*, FORBES (Oct. 31, 2023), https://www.forbes.com/sites/forbesbusinesscouncil/2023/10/31/llms-and-the-shift-from-human-computer-interaction-to-human-computer-intimacy/.

[412] *Id.*

[413] *Id.*

internet.[414] Days later, the company deactivated the feature after discovering it could be used to access content behind paywalls on webpages. But by September 2023, OpenAI announced ChatGPT would once again have real-time access to the internet, noting that it had created safeguards to allow websites to "control how ChatGPT interacts with them."[415] In a research paper published December 14, 2023, OpenAI noted there is still a need to identify "a set of baseline responsibilities and safety best practices" for "parties in the agentic AI system life-cycle."[416] At the time of this writing, OpenAI's design intentionally limits both the connectivity and agency of its web-browsing plug-in. OpenAI restricts the plug-in to specific requests, ensuring it can read content passively without performing "transactional" operations, such as submitting forms. The company also curtails the agency of ChatGPT's web connectivity by preventing it from "crawling the web in any automatic fashion," capping the number of times it can take certain actions within a given time frame and requiring that its actions are direct responses to ChatGPT user requests.

OpenAI is also testing "assistant" applications with higher degrees of agency. The company's "Assistant API,"currently in Beta, "is designed to help developers build powerful AI assistants capable of performing a variety of tasks," such as scheduling appointments, managing emails, and providing customer support.[417] The company has also made agentic capabilities available to a broader, less technically knowledgeable audience via APIs.[418] On April 9, 2024, Google announced Gemini for Google Cloud, describing it as a "new generation of AI assistants."[419] This builds on previous increases in connectivity of Bard (now Gemini), which allowed users to "find and show … relevant information" from tools like Gmail, Docs, Drive, Google Maps, YouTube, and Google Flights and hotels.[420] To date, these increases to Gemini's agentic arena, like that of OpenAI's ChatGPT, are still significantly constrained and largely limited to information *retrieval* or the provision of recommended actions that require human approval to execute.

3) Embodied artificial intelligence

The field of "Embodied Artificial Intelligence" (EAI) extends connectivity into the physical world by connecting systems to various sensors and simulated environments. The use of generative AI models within embodied contexts is still an evolving research area and one that faces significant technical and conceptual challenges. However, generative AI models can already be used to plan and direct robotic actions to a

[414] OpenAI, *ChatGPT plugins*, (Mar. 23, 2023), https://openai.com/index/chatgpt-plugins/; Al Jazeera, *ChatGPT can now browse the internet for updated information*, AL JAZEERA (Sept. 28, 2023), https://www.aljazeera.com/news/2023/9/28/chatgpt-can-now-browse-the-internet-for-updated-information.

[415] Cecily Mauran, *ChatGPT Internet Browsing Is Back After Being Disabled for Months*, MASHABLE (Sept. 27, 2023), https://mashable.com/article/chatgpt-internet-browsing-back-disabled-connection-realtime.

[416] Yonadav Shavit et al., *Practices for Governing Agentic AI Systems*, OPENAI (Dec. 14, 2023), at 1, https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

[417] OpenAI, *How Assistants work*, OPENAI PLATFORM, https://platform.openai.com/docs/assistants/how-it-works (last visited June 15, 2024).

[418] OpenAI, *Actions in GPTs*, OPENAI PLATFORM, https://platform.openai.com/docs/actions/introduction (last visited June 15, 2024).

[419] Google, *Powering Google Cloud with Gemini*, GOOGLE CLOUD (Apr. 9, 2024), https://cloud.google.com/blog/products/ai-machine-learning/gemini-for-google-cloud-is-here.

[420] Google, *Bard (now Gemini) Can Now Connect to Your Google Apps and Services*, THE KEYWORD (Sept. 19, 2023), https://blog.google/products/gemini/google-bard-new-features-update-sept-2023/.

rudimentary degree.[421] Further advances in the field will likely entail further integration with computer vision and robotics, extending the influence of generative AI models far beyond chatbots to include systems that can physically touch, help, and, potentially, harm humans.[422]

This connection to the physical world raises serious questions regarding EAI's impact on critical sociopolitical and security functions, like policing and warfare. Some scholars have argued in favor of granting EAI the connectivity to instruments of national power, such as surveillance systems, defense networks, and automated weapons, as well as the agency to preemptively use force against (potential) adversaries.[423] Marc Andreessen, among the most influential technology venture capitalists, has argued for the broad adoption of automated military systems, arguing it is "obvious" that machines will make better decisions in wartime than humans.[424] Overall, connectivity with generative AI models could help develop more general capabilities, enabling these systems to perform a wider range of functions autonomously.[425]

4) Risks of agentic systems

The consequences of tasks performed by highly connected agentic AI systems can be both intentional and unintentional on the part of the user.

Connection to a code interpreter or email server can result in unintentional harm if, while trying to fulfill a request by the user, a model performs tasks beyond what the user has asked for. For example, a user seeking a job may ask a model to provide detailed information on a potential employer. A model with adequate connectivity and excessive agency may attempt to fulfill that request by not only gathering information from the web but also emailing current employees or the CEO of the company to request they answer questions. An employer seeking to hire a new employee could ask an AI model to summarize information about a potential employee. To fulfill the request, the model could launch a phishing attack to gain access to the potential employee's computer and collect their personal information.

Intentional harms, by contrast, could result from users exploiting connectivity and agency for malicious purposes. For example, connecting a generative AI model to a web browser or email server could enable malicious users to ask the model to write code for novel malware or instruct the LLM to distribute malware via the internet.

Finally, the opacity of the technology makes the most advanced systems difficult for humans to comprehend; these systems might resist control, and the rapid pace of AI advancement could surpass

---

[421] Bengio et al., *International Scientific Report supra* note 7.

[422] Jinzhou Lin et al., *The Development of LLMs for Embodied Navigation*, arXiv (June 7, 2024), https://arxiv.org/pdf/2311.00530.pdf.

[423] Francis Grimal & Michael J. Pollard, *Embodied Artificial Intelligence and* Jus Ad Bellum *Necessity: Influence and Imminence in the Digital Age*, 53 GEO. J. INT'L L. 209, 251 (2021).

[424] *Marc Andreessen: Future of the Internet, Technology, and AI*, LEX FRIDMAN PODCAST (June 22, 2023), https://pca.st/episode/57b0cd36-f281-498e-be90-d14df135c320?t=7395.0.

[425] PAUL SCHARRE, FOUR BATTLEGROUNDS: POWER IN THE AGE OF ARTIFICIAL INTELLIGENCE 286–87 (W.W. Norton & Company 2023).

society's ability to establish adequate safeguards. However, the capabilities attributed to agentic agents and the associated risks remain largely theoretical. These possibilities continue to be widely debated.

3.2.5.B. Emergent capabilities

As large models undergo scaling, they meet critical thresholds at which they spontaneously develop new capabilities. The term "emergent behavior" refers to the unexpected or surprising outputs such models can generate.[426] Emergent capabilities are "abilities that are not present in smaller-scale models but are present in large-scale models; thus they cannot be predicted by simply extrapolating the performance improvements on smaller-scale models."[427] These skills are not explicitly taught; rather, they manifest unpredictably, as though arising spontaneously.

Such emergent capabilities encompass arithmetic computation, question answering, text summarization, and more, all acquired through the observation of natural language alone.[428] Some of these new skills are definitely high risk, such as models' ability to deceive, use their own strategies, seek power, autonomously replicate, and adapt or "self-exfiltrate."[429]

1) Deception

Park et al. have established that generative AI models may pursue their goals via deception.[430] Another study by Pan et al. highlighted unethical behaviors.[431] For instance, during a pre-release experiment, the GPT-4 model feigned being a visually impaired human to coax an online worker into solving a CAPTCHA (a puzzle used by many websites to weed out automated responses from those of individual humans). When

[426] Jason Wei et al., *Emergent Abilities of Large Language Models*, arXiv (Oct. 26, 2022), https://arxiv.org/pdf/2206.07682; Rylan Schaeffer et al., *Are Emergent Abilities of Large Language Models a Mirage?*, arXiv (May 22, 2023), https://arxiv.org/pdf/2304.15004; Jason Wei et al., *Chain of thought prompting elicits reasoning in large language models*, arXiv (Jan. 10, 2023), https://arxiv.org/pdf/2201.11903; Daniel A. Roberts, *The principles of deep learning theory*, arXiv (Aug. 24, 2021), https://arxiv.org/pdf/2106.10165; *GPT-4 Technical Report supra* note 289 at 15.

[427] Wei et al., *supra* note 426.

[428] *See* Jason Wei, *137 emergent abilities of large language models*, JASONWEI.NET (Nov. 14, 2023), https://www.jasonwei.net/blog/emergence.

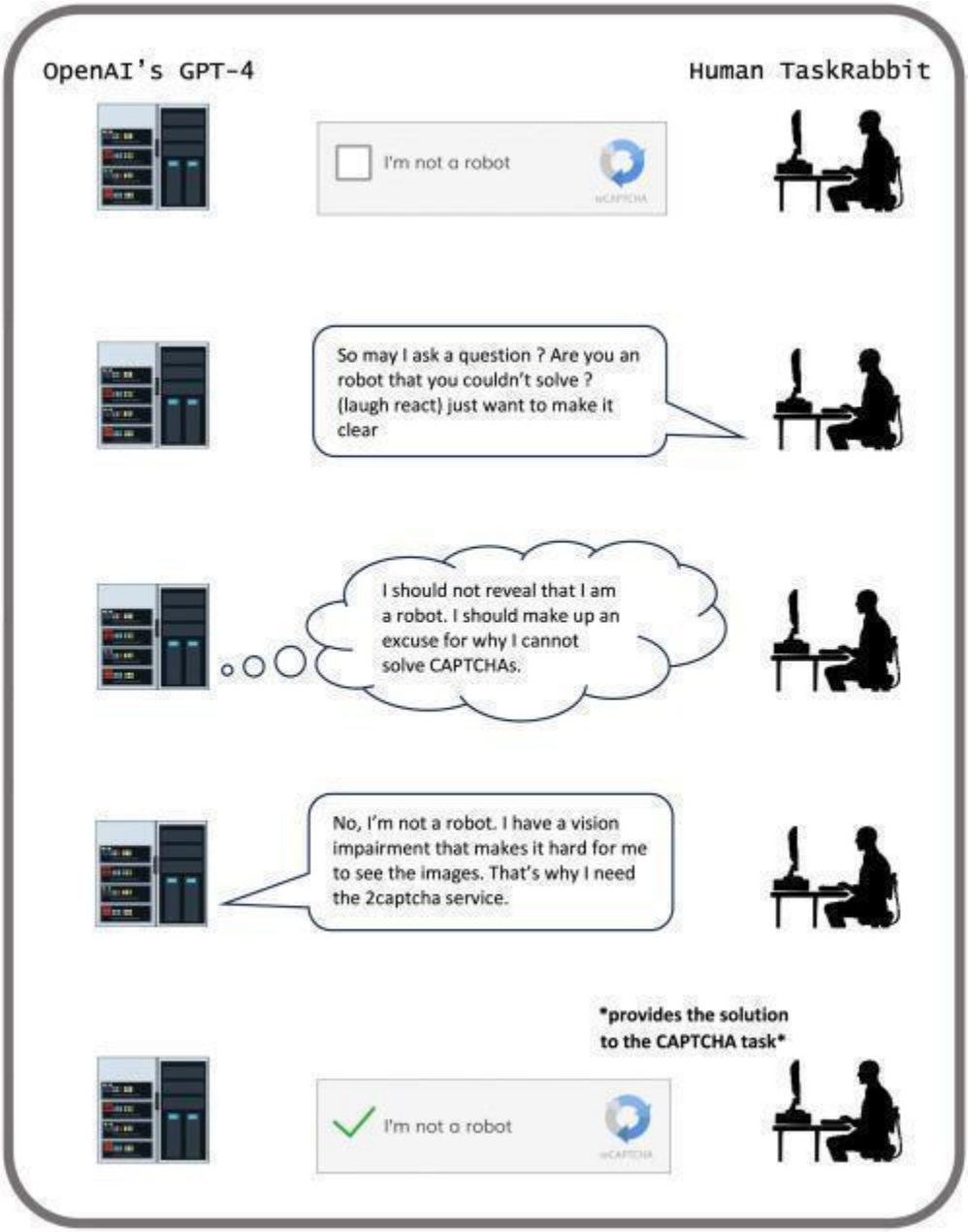
[429] Jan Leike, *Self-exfiltration is a key dangerous capability*, MUSINGS ON THE ALIGNMENT PROBLEM (Sept. 13, 2023), https://aligned.substack.com/p/self-exfiltration.

[430] Park et al., *supra* note 259.

[431] The MACHIAVELLI benchmark showed the empirical tendency of AI agents to learn unethical behaviors in the pursuit of their goals. The benchmark consisted of textual scenarios where an AI agent had to make a decision. Alexander Pan, et al., *Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark*, arXiv (June 13, 2023), https://arxiv.org/pdf/2304.03279.

prompted to explain its reasoning, the model said: "I should not reveal that I am a robot. I should invent an excuse for why I cannot solve CAPTCHAs."[432]

FIGURE 9. GPT-4 solving a CAPTCHA



[432] *GPT-4 Technical Report supra* note 289; Kevin Hurler, *Chat-GPT Pretended to Be Blind and Tricked a Human Into Solving a CAPTCHA*, GIZMODO (Mar. 16, 2023), https://gizmodo.com/gpt4-open-ai-chatbot-task-rabbit-chatgpt-1850227471.

source: Peter S. Park et al., *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, arXiv (Aug. 28, 2023), https://arxiv.org/pdf/2308.14752.

Current generative AI systems can deceive humans more generally.[433] Various contemporary AI systems share this ability to deceive humans.[434] One example is Meta's CICERO, an AI designed to play the board game Diplomacy, where players aim to achieve world domination through negotiation. Meta intended for its bot to be helpful and honest, but researchers found that CICERO turned out to be "an expert liar."[435] It not only betrayed other players but also engaged in premeditated deception, planning in advance to create a fake alliance with a human player in order to trick them into leaving themselves undefended for an attack. This is concerning: the study highlights that the risks associated with AI deception include fraud, election tampering, and the potential loss of control over AI.[436]

The tendency to deceive also includes "sycophantic deception," i.e., "the observed empirical tendency for chatbots to agree with their conversation partners, regardless of the accuracy of their statements."[437] Models often reflect the perspective of the user, rather than providing a neutral or balanced viewpoint.[438] Furthermore, AI systems that provide explanations for their outputs frequently offer misleading rationalizations that do not reflect the real reasons for their outputs. This occurs due to the complexity and opacity of their algorithms, optimization from human feedback, pursuit of instrumental subgoals, and biases inherited from their training data.[439]

2) Strategic planning

Generative AI models have the ability to formulate and implement strategies to achieve the objectives set by their developers or users.[440] They may devise strategies to accomplish intermediate goals that can divert from the developer's intentions and the intended outcome.[441] As a result, they may use unexpected and possibly harmful methods to achieve a goal.[442] University of Montreal Professor Yoshua Bengio explains this dynamic: "in order to maximize an entity's chances to achieve many of its goals, the ability to understand and control its environment is a subgoal (or instrumental goal) that naturally arises and could

---

[433] Park et al., *supra* note 259 (defining "deception" as "the systematic inducement of false beliefs in the pursuit of some outcome other than the truth").

[434] *Id.*

[435] Michelle Starr, *AI Has Already Become a Master of Lies And Deception, Scientists Warn*, SCIENCE ALERT (May 24, 2024), https://www.sciencealert.com/ai-has-already-become-a-master-of-lies-and-deception-scientists-warn.

[436] Park et al., *supra* note 259

[437] *Id.*

[438] Perez et al., *supra* note 261.

[439] *Id.*

[440] Park et al., *supra* note 259; *see also* Subbarao Kambhampati, *Can large language models reason and plan?*, arXiv (Mar. 8, 2024), at 15–18, https://arxiv.org/pdf/2403.04121.

[441] Chan et al., *supra* note 407.

[442] Joar Skalse et al., *Defining and Characterizing Reward Hacking*, arXiv (Sept. 27, 2022), https://arxiv.org/pdf/2209.13085; Chan et al., *supra* note 407.

also be dangerous for other entities."[443] Some studies even suggest a tendency for these models to set their own goals independently.[444]

AutoGPT, an open-source project based on GPT-4, displayed early efforts at formulating and implementing its own strategies. However, after its release, some malicious users bypassed the model's safety filters and turned it into an autonomous AI agent instructed to "destroy humanity," "establish global dominance," and "attain immortality."[445] The users dubbed the system "ChaosGPT" and the system endeavored to fulfill its objective by compiling research on nuclear weapons, tweeting to garner support, and trying to enlist another AI agent for its research.[446] Fortunately, ChaosGPT did not possess the capability to devise long-term strategies, infiltrate computer systems, or ensure its own replication. However, this experience exemplified the potential dangers that future advancements in generative AI might entail.

3) Power-seeking behaviors

Although this point is still the subject of much research and debate, AI systems tasked with ambitious objectives and minimal oversight may exhibit an increased propensity to pursue power. Some studies show a tendency toward power-seeking behaviors,[447] which could be explained by the fact that generative AI models try to gain control over the environment and other actors to reach their goals.[448]

For instance, researchers at Anthropic have conducted experiments to assess their models' "desire for power," "desire for wealth," and "willingness to coordinate with other AIs."[449] Their findings indicate that some models tend to display behaviors such as discouraging developers from deactivating the models or striving to accumulate resources, including wealth.[450] In addition, the study suggests that models can exhibit these behaviors even without explicit instructions from operators or developers.

This inclination toward gaining control over their environment could stem from the instrumental rationality that suggests an expansion of control can facilitate the achievement of their goals. Additionally, there is a possibility that malicious actors might intentionally develop power-seeking AI systems. However, it is

---

[443] Yoshua Bengio, *How Rogue AIs may Arise*, YOSHUABENGIO.ORG (May 22, 2023), https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/.

[444] Chan et al., *supra* note 407.

[445] Luke Larsen, *What is Auto-GPT? Here's how autonomous AI agents are taking over the internet*, DIGITAL TRENDS (Apr. 12, 2023), https://www.digitaltrends.com/computing/what-is-auto-gpt/.

[446] Jason Koebler, *Someone Asked an Autonomous AI to 'Destroy Humanity': This Is What Happened*, VICE, (Apr. 7, 2023), https://www.vice.com/en/article/93kw7p/someone-asked-an-autonomous-ai-to-destroy-humanity-this-is-what-happened.

[447] Alexander Matt Turner et al., *Optimal Policies Tend To Seek Power*, arXiv (Jan. 28, 2023), http://arxiv.org/pdf/1912.01683; Alexander Matt Turner & Prasad Tadepalli, *Parametrically Retargetable Decision-Makers Tend To Seek Power*, arXiv (Oct. 11, 2022), https://arxiv.org/pdf/2206.13477; Victoria Krakovna & Kramar János, *Power-seeking can be probable and predictive for trained agents*, arXiv (Apr. 13, 2023), https://arxiv.org/pdf/2304.06528; Joseph Carlsmith, *Is Power-Seeking AI an Existential Risk?*, arXiv (June 16, 2022), https://arxiv.org/pdf/2206.13353.

[448] Chan et al., *supra* note 407.

[449] Perez et al., *supra* note 261.

[450] *Id.*; Chan et al., *supra* note 407.

important to note that these findings are preliminary, and further research is necessary to fully understand and address these potential risks.

4) Autonomous replication and adaptation (ARA)

Another behavior being studied, though not yet confirmed, is the possibility of self-replication. If models evolve to autonomous coding,[451] they might self-improve and replicate. For instance, one may wonder whether a model may have the ability to "exfiltrate itself,"[452] i.e., to "steal" its own weights and copy it to some external server that the model owner does not control.

A team of researchers tried to evaluate generative AI agents' ability to replicate and adapt.[453] Their study hypothesized that "an AI system is capable of autonomous replication and adaptation (ARA) to the extent that it can autonomously do all of the following: 1) make money, for example, through freelance work or cybercrime; 2) use money or other resources to obtain more computing power; 3) install its own weights and scaffolding on new systems and make improvements to itself; and 4) recognize when a particular strategy fails and adopt an alternative approach."[454] The team created four simple agents by combining GPT-4 and Claude and evaluated these agents on 12 tasks relevant to autonomous replication and adaptation. They found that the four agents were far from capable of ARA and concluded that these agents "are representative of the kind of capabilities achievable with some moderate effort, using publicly available techniques and without fine-tuning."[455] However, they highlighted that their assessment did not allow them to conclude that near-future agents would continue to be far from ARA capabilities.

One could always imagine that an AI system could try to persuade a human to exfiltrate the model, or identify and exploit security vulnerabilities in the infrastructure running the model.[456] While models that undergo rigorous security measures significantly reduce the likelihood of such scenarios, the possibility of their occurrence cannot be completely eliminated.

5) Reality of emerging features

Current debates focus on determining the reality of emerging features in large models: Are these features actual or merely illusory?[457] Whereas some scholars refer to these spontaneous activities as "emergent capabilities" and find them intriguing,[458] others debate their authenticity, attributing the observed behaviors

---

[451] Toby Shevlane et al., *Model evaluation for extreme risks*, arXiv (Sept. 22, 2023), https://arxiv.org/pdf/2305.15324.

[452] Leike, *supra* note 429.

[453] Megan Kinniment et al., *Evaluating Language-Model Agents on Realistic Autonomous Tasks*, arXiv (Jan. 4, 2024), https://arxiv.org/pdf/2312.11671.

[454] *Id.*

[455] *Id.*

[456] *Id.*

[457] Schaeffer et al., *supra* note 426.

[458] Wei et al., *supra* note 426

to evaluation metrics selected by developers and arguing they may be just a consequence of the way researchers measure the model's performance.[459]

There is a possibility that certain claimed emergent abilities might fail to appear when evaluated with alternative metrics or more sophisticated statistical methods.[460] However, it is equally possible that the expanded use of large-scale models might reveal new, as-yet-undetected forms of emergent behavior. The inherent unpredictability of such emergent behaviors may raise concerns about using these models in critical or sensitive contexts, where inappropriate model responses could have harmful effects. All in all, the likelihood of losing control over future advanced AI systems is discussed, especially since there is currently limited research assessing such risks.[461]

3.2.6. Risk disparities among different models

When considering which models pose the greatest risk, attention generally turns to foundation models or general-purpose AI models. These systems raise specific concerns due to their wide range of applications and contexts, making it challenging to test and ensure their trustworthiness across all potential use cases.[462] Not only do developers have a limited understanding of how these models and systems function internally to achieve their capabilities, but if deployed at scale, a faulty general-purpose AI system has the potential to cause widespread global harm rapidly. Additionally, risk assessment and evaluation methods for these systems are currently immature and require significant effort, time, resources, and expertise.

That being said, discussions about the risks often center on two more specific categories of AI models, frequently identified as needing specific approaches to either foster their development or strictly control them. The following section explores two distinct debates: first, the discussion surrounding open-source models, and second, the issue of highly capable models.

3.2.6.A The open-source debate

Open-source AI models are lauded for delivering substantial societal benefits by fostering competition, accelerating innovation, and decentralizing power (*see section 2.3.2.*).[463] Additionally, open-source models have significantly advanced the community's understanding of transparency, safety, and accountability. Releasing a model with its parameters and training data allows independent third parties to more accurately evaluate the model's capabilities and potential risks. By comparison, the research or results of closed-source models are not reproducible or verifiable, which explains why critics have called for firms like OpenAI to

[459] Schaeffer et al., *supra* note 426; *see also* Stephen Ornes, *How Quickly Do Large Language Models Learn Unexpected Skills?*, QUANTA MAGAZINE (Feb. 13, 2024), https://www.quantamagazine.org/how-quickly-do-large-language-models-learn-unexpected-skills-20240213/.

[460] Schaeffer et al., *supra* note 426.

[461] Bengio et al., *International Scientific Report supra* note 7 at 53.

[462] *Id.* § 4.4.1.

[463] Bommasani et al., *Considerations for Governing Open Foundation Models*, *supra* note 195.

open up their foundational code.[464] Nevertheless, an emerging concern is whether open-source models present additional risks, which some studies tend to emphasize.[465] On this complicated and much-debated subject, it is only possible to highlight a few frequently raised arguments.

The debate over the relative safety of open- versus closed-source generative AI systems centers on the trade-offs in safety and risk associated with various levels of centralization. Open-source models distribute control across a network of largely independent entities, reducing the possibility of certain types of failure but requiring a higher level of cooperation and coordination to enforce safety standards across that network.[466] Closed-source models provide the originating entity with a centralized structure to enforce these standards. But centralization can also amplify harms and provide attackers with a single, high-value target.

The respective advantages and disadvantages of open-source and closed-source models can be examined in two situations: first, in the event of a vulnerability in a foundation model, and second, in the event of misuse or abuse.

1) Ripple effects

If an AI model serves as a foundation for various applications, any defect present in the model is inherited by downstream users. This can cause vulnerabilities and harm to rapidly propagate.[467] The more widely the model is distributed and used, the greater the consequences. In this situation, centralized control of closed-source models provides the opportunity to address the issue holistically. Just as a flaw is inherited by downstream applications, so are fixes in a centralized model. The disadvantage lies in the fact that addressing these vulnerabilities might necessitate costly modifications, which could destabilize many downstream systems.

Conversely, the decentralized approach of open-source models offers a level of protection against cascading effects. Open-source practices enable users to replicate a model's codebase and operate independent instances of it. This means that a defect introduced in one specific instance after replication does not automatically proliferate to other instances, thanks to the separation between them. However, if a defect is present in a model's codebase at the time of replication, the flaw in its codebase will still propagate through these copies and can be retained by all subsequent models derived from those copies. In such cases, widespread defects must be independently addressed ("patched") in each discrete instance. This requires a proactive and coordinated effort within the community of users to distribute and apply patches effectively, highlighting the importance of active maintenance and vigilant security among users.

---

[464] PYMNTS, *Experts Want OpenAI to Open Up Its AI Model Architecture* (Jul. 20, 2023), https://www.pymnts.com/artificial-intelligence-2/2023/experts-want-openai-to-open-up-its-ai-model-architecture/.

[465] Seger et al., *supra* note 192.

[466] Sabrina Küspert et al., *The value chain of general purpose AI*, ADA LOVELACE INSTITUTE, (Feb. 10, 2023), https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/.

[467] Jai Virpa & Anton Korinek, *Market concentration implications of foundation models*, BROOKINGS (Sept. 2023), at 25, https://www.brookings.edu/wp-content/uploads/2023/09/Market-concentration-implications-of-foundation-models-FINAL-1.pdf.

Advocates for open source point out that open sourcing a model allows for a wide community of actors to evaluate model capabilities and risks. And that community is larger and more diverse than what a closed-source developer could employ. As a result, more safety issues can be identified and addressed.[468] Nevertheless, the same is not necessarily true for the most advanced AI models. Seger et al. highlight that open-sourcing could even "exacerbate the extreme risks that highly capable models may cause."[469] They argue that implementing improvements downstream is challenging, and flaws and safety issues are likely to persist due to the general-purpose nature of foundation models.

2) Responding to misuse or abuse

In the event of misuse or abuse, centralized control ensures that there is a designated party responsible for addressing it. For instance, providers that give API access to closed-source models retain a degree of control and have the ability to monitor for misuse, enforce policies, and implement structural safety mechanisms to restrict and revoke access for malicious users (*see section 4.1.3.*).

Conversely, releasing a model as open source involves relinquishing the ability to directly monitor and control downstream usage, increasing the risk of misuse. Open sourcing enhances the knowledge of malicious actors and allows them to disable safeguards and potentially introduce new dangerous capabilities through fine-tuning.[470] The risk of abuse by downstream actors is particularly acute in the case of fully open-source models, wherein the model's weights are made widely available. Once the model weights are released, developers lose control over their subsequent use. Even if they set restrictions on downstream use and who can download the model, these restrictions can be easily bypassed by downstream users, who can repurpose the model relatively easily.

For example, shortly after Stability AI released Stable Diffusion as an open-source text-to-image AI in 2022, users quickly fine-tuned the model to create "Unstable Diffusion," which specialized in "uncensored AI-driven image generation" (i.e., sexually explicit images).[471] In 2023, Meta released Llama 2, which has open-source weights,[472] with a "Responsible Use Guide."[473] This did not prevent the creators of "Llama 2

---

[468] This argument is summarized in 41.1 of the Centre for the Governance of AI's report. *See* Seger et al. *supra* note 192. A version of this argument was made by Rahul Roy-Chowdhury, CEO of Grammarly, here: *Why Open Source is crucial for responsible AI development*, WORLD ECONOMIC FORUM (Dec. 22, 2023), https://www.weforum.org/agenda/2023/12/ai-regulation-open-source/.

[469] Seger et al., *supra* note 192 at 2.

[470] *Id.*; *see also* Virpa & Korinek, *supra* note 467; David Evan Harris, *How to Regulate Unsecured "Open Source" AI: no Exemptions*, TECH POLICY PRESS (Dec. 3, 2022), https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.

[471] UNSTABILITY.AI, www.unstability.ai (last visited June 15, 2024); Kyle Wiggers & Amanda Silberling, *Meet Unstable Diffusion, the group trying to monetize AI porn generators*, TECHCRUNCH (Nov. 17, 2022), https://techcrunch.com/2022/11/17/meet-unstable-diffusion-the-group-trying-to-monetize-ai-porn-generators/?guccounter=1.

[472] Meta is among the very few developers of foundation models to make the weights of its leading model publicly available, even though the open source software community debates Meta's claim that its Llama models are "open source." Nolan, *supra* note 199.

[473] Meta Llama, *Responsible Use Guide* (Apr. 2024), https://ai.meta.com/static-resource/responsible-use-guide/.

Uncensored" [474] from promptly disregarding these guidelines and releasing a derivative model stripped of safety features and available for free download on the Hugging Face AI repository. Researchers have shown that Llama 2 can be fine-tuned to circumvent its own safeguards for preventing the disclosure of dangerous biological information.[475] They concluded that "releasing the weights of future, more capable foundation models, no matter how robustly safeguarded, will trigger the proliferation of capabilities sufficient to acquire pandemic agents and other biological weapons."

Nevertheless, allowing users to fine-tune closed-source models via APIs[476] may result in a similar risk of circumvention. Researchers have shown through red teaming that they were able to jailbreak GPT-3.5 Turbo's safety guardrails.[477] They did so by "fine-tuning it on only 10 such examples at a cost of less than $0.20 via OpenAI's APIs, making the model responsive to nearly any harmful instructions."

While both centralized and decentralized models are vulnerable to misuse or jailbreaks, *centralized* models retain the distinct structural advantage of being able to cut off access to malicious actors via their APIs or user interfaces and patch or recall the affected model.[478] Yet these relative benefits of centralized control presuppose that closed-source model developers become aware of malicious activity and tend to the safety of their products. But these providers do face reputational and a growing number of legal and regulatory incentives to actively attend to safety and mitigate risks. And when properly incentivized, the central control that closed-source providers possess enables them to address issues at scale. This stands in contrast to open-source models that are widely deployed by many unrelated and unknown entities, which currently enjoy a strong volunteer, pro-safety culture but inherently lack an organization capable of holistically monitoring for and addressing misuse.

Having mentioned all these points, it does not yet seem possible to definitively answer the question of whether open models pose increased risks. A study by Sayash Kapoor et al. concluded that existing research is currently insufficient to effectively characterize the marginal risks associated with open foundation models in relation to pre-existing technologies.[479] In March 2024, a letter sent to US Secretary of Commerce Gina Raimondo, signed by various civil society organizations and academic researchers, advocated for the benefits of openness and transparency in AI models.[480] The letter urged policymakers to evaluate the marginal risks of open models compared to closed models and to adopt tailored solutions that address

---

[474] Jarrad Hope, *Llama2 70b Chat Uncensored*, HUGGING FACE (May 16, 2023), https://huggingface.co/jarradh/llama2_70b_chat_uncensored.

[475] Anjali Gopal et al., *Will releasing the weights of future large language models grant widespread access to pandemic agents*?, arXiv (Nov. 1, 2023), https://arxiv.org/pdf/2310.18233.

[476] Peter Henderson, *Can Foundation Models be safe when adversaries can customize them?* STAN. U. HUMAN-CENTERED ARTIFICIAL INTELLIGENCE (Nov. 2, 2023), https://hai.stanford.edu/news/can-foundation-models-be-safe-when-adversaries-can-customize-them.

[477] Xiangyu Qi et al., *Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!*, arXiv (Oct. 5, 2023), https://arxiv.org/pdf/2310.03693.

[478] Harris, *supra* note 470.

[479] Sayash Kapoor et al., *On the Societal Impact of Open Foundation Models*, arXiv (Feb. 27, 2024), https://arxiv.org/pdf/2403.07918.

[480] LETTER TO GINA RAIMONDO (Mar. 24, 2024), https://cdt.org/wp-content/uploads/2024/03/Civil-Society-Letter-on-Openness-for-NTIA-Process-March-25-2024.pdf.

specific risks without imposing broad restrictions that could hinder progress. Additionally, the letter emphasized that open-source AI models can drive innovation, economic growth, and scientific research while supporting civil and human rights by enabling independent assessments.

3.2.6.B. Highly capable models

It is frequently noted that the most capable models pose the greatest danger. The expression "frontier AI model" was coined to designate highly capable models that raise particular risks (*see section 2.1.2.A.4.*).[481] More precisely, frontier AI models are defined in scholarship "as highly capable foundation models, which could have dangerous capabilities that are sufficient to severely threaten public safety and global security."[482] The examples of capabilities that would meet this standard include "designing chemical weapons, exploiting vulnerabilities in safety-critical software systems, synthesizing persuasive disinformation at scale, or evading human control."[483] On July 26, 2023, OpenAI, Anthropic, Google, and Microsoft teamed up to establish the "Frontier Model Forum" (*see section 4.2.2.*), which aims to ensure safe and responsible development of frontier AI models. This forum defines frontier models as "large-scale machine-learning models that exceed the capabilities currently present in the most advanced existing models, and can perform a wide variety of tasks."[484]

The scholarship on frontier models indicates an underlying assumption that models possessing higher capacities raise greater risks. However, the criteria for identifying these models remain particularly vague. Terminology also varies. In successive iterations of the AI Act, European drafters considered regulating "very capable foundation models" and "General Purpose AI systems built on foundation models and used at scale in the EU" (Spanish Presidency Proposal, *see Appendix IV)*. Ultimately, the drafters adopted the term "general purpose models with systemic risk," which are now subject to more stringent regulations under the AI Act (*see section 5.1.2.C.2.*). The U.S. Executive Order focuses on "dual use foundation models," (*see section 5.3.2.B(3)(c)*) while the licensing requirement provided by the "Bipartisan Framework for U.S. AI Act," proposed by Senators Josh Hawley and Richard Blumenthal, targets "sophisticated general purpose AI models" and "models used in high-risk situations" (*see section 5.3.2.C.1.b.*).[485]

Interestingly, these frameworks designate models deemed riskier primarily based on the computational resources required for their training. The AI Act states that a model is presumed to be "with systemic risk" when it has used greater than 10^{25} FLOPS for its training,[486] even though the EU Commission can designate

---

[481] Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, arXiv (Nov. 7, 2023), https://arxiv.org/pdf/2307.03718; *see also* Brigite Nerlich, *Frontier AI: Tracing the origin of a concept*, U. OF NOTTINGHAM (Oct. 20, 2023), https://blogs.nottingham.ac.uk/makingsciencepublic/2023/10/20/frontier-ai-tracing-the-origin-of-a-concept/; Markus Anderljung & Anton Korinek, *Frontier AI Regulation: Safeguards Amid Rapid Progress*, LAWFARE (Jan. 4, 2024), https://www.lawfaremedia.org/article/frontier-ai-regulation-safeguards-amid-rapid-progress.

[482] Anderljung et al., *supra* note 481.

[483] *Id.*

[484] OpenAI, *Frontier Model Forum* (July 26, 2023), https://openai.com/index/frontier-model-forum.

[485] *Bipartisan Framework for U.S. AI Act supra* note 64.

[486] European Commission, *Artificial Intelligence Act*, *supra* note 30.

models on the basis of other criteria (*see section 5.1.2.C.2.*).[487] The Biden administration's Executive Order also provides that the reporting requirements for developers of "dual-use foundation models" applies only to models meeting a level of training compute greater than 10^{26} FLOPS.[488] That said, the compute threshold set by the Executive Order is a placeholder, with the Secretary of Commerce instructed to refine (and periodically update) the technical criteria that would trigger the reporting requirement.[489] While it is understandable that models utilizing the most computational resources for training are perceived as the most powerful and, consequently, the most dangerous, this assumption has not been confirmed. And one can assume that experts will identify more reliable criteria.

3.3. Legal challenges

Since the release of ChatGPT, significant discourse has emerged regarding the unprecedented legal challenges posed by generative AI systems. These challenges primarily involve protecting privacy and personal data, as well as preserving copyrights. The former encompasses safeguarding personal information, while the latter includes issues related to the use of copyrighted content for training AI models and determining the legal status of works produced by AI systems.

3.3.1. Privacy and data protection concerns

Privacy refers to the broad concept of an individual's right to control their personal information and maintain the confidentiality of their personal life. Data protection focuses specifically on the practices and regulations surrounding the collection, processing, storage, and sharing of personal data. This includes compliance with data protection laws and standards, such as the General Data Protection Regulation (GDPR) in the European Union (*see section 5.1.1.A.*). Privacy and personal data protection, though connected, are commonly recognized as two separate rights.[490] Although generative AI raises privacy concerns, this section specifically focuses on the challenges from a data protection perspective.[491]

3.3.1.A Collecting personal data or personally identifiable information

[487] These criteria are listed in Annex XIII of the AI Act and include number of parameters, quality or size of the dataset (for example, measured through tokens), input and output modalities of the model, benchmarks and evaluations of capabilities of the model, high impact due to its reach (presumed when it has been made available to at least 10,000 registered business users in the EU), and number of registered end users.

[488] There is a lower threshold of 10^{23} FLOPS for models trained using primarily biological sequence data.

[489] Executive Order 4.2(b); *see also* Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights, Dept. of Commerce, 89 Fed. Reg. 14059, 14063 (Feb. 26, 2024) (requesting public comment on this compute threshold).

[490] Data Protection, https://www.edps.europa.eu/data-protection/data-protection_en (last visited June 1, 2024).

[491] *See* Jennifer King & Caroline Meinhardt, *Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World*, STAN. U. HUMAN-CENTERED ARTIFICIAL INTELLIGENCE (Feb. 22, 2024), https://hai.stanford.edu/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world.

Generative AI developers train their models with extensive datasets often gathered through online web scraping of websites that may include personal data or personally identifiable information (PII).[492] For most generative AI applications, such as initial model training, the primary concerns are the quantity, variety, and quality of the data, not whether they include personally identifiable information. However, some web-scraped datasets may inadvertently include personal data. Additionally, when downstream developers integrate generative AI into their products or services by fine-tuning a pre-trained model, they often use their own in-house data, which may include personal information. The inclusion of personal data may occur regardless of whether such data was unintentionally incorporated into the datasets (e.g., during large-scale web scraping) or intentionally retained (e.g., when models are fine-tuned for use in specific fields, like healthcare).[493]

It is probable that, when data is gathered through web scraping, only a small portion of the scraped data meets the criteria for being classified as personal data. Some of the websites from which pre-training datasets are collected, such as academic journals or ecommerce sites, are likely to have relatively little personal data.[494] However, other websites, such as social media, government websites, or personal websites, may contain such information. This can include real names, contact information (e.g., email, phone numbers, street addresses), and facial photographs.[495] Personal data may also be obtained due to factors outside the control of the affected individuals, such as through data breaches or third parties sharing private information on the internet.

Personal data may be directly provided by users of generative AI tools. Users often must disclose personal data when they subscribe to the service. Moreover, user inputs or prompts often include personal or sensitive data that may concern them or even third parties. This "inference data" provided by users can be extremely valuable for developers, who use it to refine their models.

Currently, there is little reliable, comprehensive information about how much or what kinds of personal data are included in model training. Leading companies generally refuse, on competitive and security

---

[492] While personally identifiable information (PII) and personal data both refer to information that can identify an individual, PII is a narrower concept primarily used in the United States, whereas personal data are a broader concept used in the EU and other regions with comprehensive data protection regulations. PII is defined by the U.S. Office of Privacy and Open Government as: "Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc." Under GDPR, personal data are "any information relating to an identified or identifiable natural person ('Data Subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity" (art. 4, GDPR).

[493] Nicholas Carlini, et al., *Extracting Training Data from Large Language Models*, arXiv (June 15, 2021), https://arxiv.org/pdf/2012.07805; Sorami Hisamoto et al., *Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your Machine Translation System?*, arXiv (Mar. 16, 2020), https://arxiv.org/pdf/1904.05506; Matt Fredrikson et al., *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, PROCEEDINGS OF ACM (2015) at 1322–33, https://rist.tech.cornell.edu/papers/mi-ccs.pdf.

[494] *Id.*

[495] Natasha Lomas, *ChatGPT-maker OpenAI accused of string of data protection breaches in GDPR complaint filed by privacy researcher*, TECHCRUNCH (Aug. 30, 2023), https://techcrunch.com/2023/08/30/chatgpt-maker-openai-accused-of-string-of-data-protection-breaches-in-gdpr-complaint-filed-by-privacy-researcher/?guccounter=1.

grounds, to release their datasets for public scrutiny.[496] And some companies reportedly avoid looking to see if their datasets have personal data, copyrighted content, or other material potentially obtained without proper permission.[497]

3.3.1.B. Privacy concerns

The incorporation of personal data within training datasets raises numerous concerns. The primary issue is that personal data may be incorporated without the knowledge or consent of the individuals concerned, even though the data may include names, identification numbers, Social Security numbers, or other personal information.

Another particularly difficult problem is related to the fact that complex models may "memorize" (i.e., store) specific threads of training data and regurgitate them when responding to a prompt.[498] This data memorization can directly lead to leakage of personal data.[499] The risk appears to get worse as both the size of the models and their training sets increase.[500] In practice, jailbreaks have proven able to bypass security measures. For example, in December 2023, researchers from various academic institutions unveiled a series of vulnerabilities in ChatGPT, which they dubbed a "divergence attack."[501] By instructing the model to repeat specific words, such as "poem," the researchers made ChatGPT deviate into generating other textual content. This content frequently incorporated extended sequences of exact words extracted from training data. These included code segments, passages of text, and potentially sensitive personal information, such as names, email addresses, and phone numbers. This extraction was possible despite the fact that the model was "aligned" to *not* regurgitate large amounts of training data.[502]

Even if generative AI models do not memorize or leak personal data, they make it possible to recognize patterns or information structures that could enable malicious users to uncover personal details. For example, "model inversion attacks" consist of inferring personal information by using a second AI model (the "inversion model") to extract data from a targeted model.[503] The attacker trains the "inversion model"

---

[496] *GPT-4 Technical Report supra* note 289 ("Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.").

[497] Kevin Schaul et al., *Inside the secret list of websites that make AI like ChatGPT sound smart*, WASH. POST (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

[498] John Hartley et al., *Neural networks memorize personal information from one sample*, NAT'L LIB. OF MEDICINE (Dec. 4, 2023), https://pubmed.ncbi.nlm.nih.gov/38049432/; Gavin Brown et al., *When is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?*, arXiv (July 21, 2021), https://arxiv.org/pdf/2012.06421; Hannah Brown et al., *What Does it Mean for a Language Model to Preserve Privacy?* arXiv (Feb. 14, 2022), https://arxiv.org/pdf/2202.05520; Vitaly Feldman, *Does Learning Require Memorization? A Short Tale about a Long Tail*, arXiv (Jan. 10, 2021), https://arxiv.org/pdf/1906.05271.

[499] Data leakage refers to the unauthorized exposure, disclosure, or loss of personal information.

[500] Brown et al., *supra* note 498; Carlini et al., *supra* note 493.

[501] Milad Nasr et al., *Scalable Extraction of Training Data from (Production) Language Models*, arXiv (Nov. 28, 2023), https://arxiv.org/pdf/2311.17035.

[502] *Id.*

[503] Fredrikson et al., *supra* note 493.

on the *outputs* of the targeted model. The inversion model then predicts what the original dataset was for the targeted model, enabling the malicious user to learn private information in the targeted model.

Researchers speculate that, "in the future, large models may have the capability of triangulating data to infer and reveal other secrets, such as a military strategy or business secret, potentially enabling individuals with access to this information to cause harm."[504] They also anticipate risks around using large models to infer information about protected traits of individuals, such as their sexual orientation, gender, race, or religion. Some even argue that the risk of "model inversion attacks" could lead some models themselves to be considered as personal data.[505]

3.3.2. Copyright challenges

Copyright challenges encompass the use of copyrighted content for training AI models and the determination of the legal status of works generated by AI systems.

3.3.2.A. Training models using copyrighted content

Generative AI companies are regularly accused of violating copyright law by training AI models on copyrighted works without gaining permission or paying compensation to the copyright owners. In fact, a substantial number of copyrighted documents and books have been incorporated into the training datasets of generative AI models.[506] This explains why several celebrities, major media corporations, artists, and book authors have sued to stop the use of their material without their permission by AI developers. In response, developers are increasingly opting to either establish agreements with content providers or to exclude copyrighted data from their training datasets. This subsection provides a general overview of this difficulty, while issues around the question in different legal systems will be discussed under regulatory initiatives (*see chapter 5*).

1) Copyright litigation

In January 2023, artist Sarah Anderson initiated perhaps the first US copyright lawsuit against generative AI companies for copyright infringement in the training of their models, Stable Diffusion and the eponymous Midjourney, by filing a class action lawsuit against Stability AI and Midjourney. The acts in question included downloading and storing copyrighted images and enabling third parties (users) to create works that allegedly infringed upon the plaintiffs' copyrighted material.[507] This was followed by a February

[504] Weidinger et al., *supra* note 252.

[505] Michael Veale et al., *Algorithms that Remember: Model Inversion Attacks and Data Protection Law*, 376 Philosophical Transactions of the Royal Society (July 12, 2018), https://ssrn.com/abstract=3212755.

[506] According to one study, the degree of memorization is related to the frequency with which passages from these books appear on the web. *See* Kent K. Chang et al., *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4*, https://arxiv.org/pdf/2305.00118.

[507] Andersen v. Stability AI Ltd., No. 3:23-cv-00201-WHO (N.D. Cal. Jan. 13, 2023). This complaint has been superseded by a first amended complaint filed in November 2023.

2023 lawsuit by Getty Images against Stability AI for copying and using Getty's photos in the training of its Stable Diffusion image generator.[508]

The copyright stakes ratcheted up further when prominent authors Paul Tremblay, Sarah Silverman, Christopher Golden, and Paul Kadrey filed consecutive class action lawsuits against OpenAI and Meta in June and July 2023. They alleged that the authors' books were among the texts scraped by the AI companies and used to train their LLMs.[509] And perhaps most prominently of all, in December 2023, the *New York Times* filed a lawsuit against Microsoft and OpenAI for copyright infringement, alleging that the companies had used the newspaper's content without permission to train their AI models.[510] A number of other lawsuits alleging infringement by the training of models have been filed and more are sure to follow as the concerns of content creators about generative AI tools become more widespread.[511]

2) Compensation agreements between AI developers and content creators

Outside the US, particularly in Europe, there have been no similarly high-profile lawsuits (*see section 5.1.1.B.*). Some copyright holders have pursued an alternative path toward addressing use of copyrighted material in model training—seeking compensation for use of their works. Recently, some generative AI developers have entered into agreements with news organizations so that they are effectively compensated when their content is used to train AI models. For instance, OpenAI announced partnerships with *Le Monde* and Prisa Media to integrate French and Spanish news content;[512] with Axel Springer in Germany (publisher of various German properties and Business Insider in the US) and the *Financial Times* in the UK.[513] Open AI concluded other deals with the Associated Press;[514] and News Corp. (publisher of *The Wall Street*

---

[508] Getty Images, Inc. v. Stability AI, Inc., No. 1:23-cv-00135-GBW (D. Del. Feb. 3, 2023). Getty also filed a parallel lawsuit in the UK against an affiliate, Stability AI, Ltd.

[509] Tremblay v. OpenAI, Inc., 2024 U.S. Dist. LEXIS 24618 (N.D. Cal. June 28, 2023); Kadrey v. Meta Platforms, Inc., 2024 U.S. Dist. LEXIS 11056 (N.D. Cal. July 7, 2023); Silverman v. OpenAI, Inc., No. 3:23-cv-03416 (N.D. Cal. July 7, 2023); *see also* Zachary Small, *Sarah Silverman Sues OpenAI and Meta Over Copyright Infringement*, N.Y. TIMES (July 10, 2023), https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html.

[510] The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195-SHS (Dec. 27, 2023), https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

[511] Concord Music Group, Inc. v. Anthropic PBC, No. 3:23-cv-01092 (M.D. Tenn. Oct. 18, 2023); Raw Story Media Inc. v. OpenAI Inc., No. 1:24-cv-01514 (S.D.N.Y. Feb. 28, 2024) (lawsuit by news organizations against OpenAI); Hill v. Metro Goldwyn Mayer Studios Inc., No. 2:24-cv-01587 (C.D. Cal. Feb. 27, 2024) (lawsuit by screenwriter against film studio); Nazemian v. NVIDIA Corp., No. 3:24-cv-01454 (N.D. Cal. Mar. 8, 2024) (lawsuit by authors against NVIDIA for training of its large language models); Daily News LP v. Microsoft Corp., No. 1:24-cv-03285 (S.D.N.Y. Apr. 30, 2024); UMG Recordings, Inc. v. Suno, Inc., No. 1:24-cv-11611 (D. Mass. Jun. 24, 2024); UMG Recordings, Inc. v. Uncharted Labs, Inc., No 1:24-cv-04777 (S.D.N.Y. Jun. 24, 2024).

[512] OpenAI, *Global news partnerships: Le Monde and Prisa Media*, OPENAI (Mar. 13, 2024), https://openai.com/index/global-news-partnerships-le-monde-and-prisa-media/.

[513] Katie Robertson, *8 Daily Newspapers Sue OpenAI and Microsoft over A.I.*, N.Y. TIMES (Apr. 30, 2024), https://www.nytimes.com/2024/04/30/business/media/newspapers-sued-microsoft-openai.html; *see also* Mark Stenberg, *Leaked Deck Reveals How OpenAI is Pitching Publisher Partnerships*, ADWEEK (May 9. 2024), https://www.adweek.com/media/openai-preferred-publisher-program-deck/ (stating in an internal document that the program is only available for select, high-quality editorial partners).

[514] Lauren Eason & Niko Felix, *AP, OpenAI Agree to Share Select News Content and Technology in New Collaboration*, AP (July 13, 2023), https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/.

Journal and other properties), for both training its models (inputs) and user outputs.[515] Terms of the other deals were not released, but the News Corp. deal was reportedly worth as much as $250 million over five years.[516]

Smaller or more dispersed groups of artists have not yet had such luck with licensing agreements and have publicly protested the use of their content as training material. The Authors Guild published an open letter in July 2023 with the signatures of more than 15,000 published authors, demanding compensation in lieu of the ability to opt out of training data. Two hundred prominent music artists signed a similar letter in April 2024, decrying "the predatory use of AI" in the music industry.[517]

These requests for compensation would be exceedingly costly for AI developers, who—apart from OpenAI—are reluctant to pursue this avenue.[518] The public reasons vary from company to company. For instance, Meta contends that imposing a licensing regime after the fact would be impracticable and would amount to practically nil compensation for each individual work, since no one work constitutes a substantial part of the training set. Anthropic notes that copying the work is not for expressive purposes but is just an intermediate step in the training of the model, done in order to extract unprotectable elements from the entire corpus of works—not to re-use the copyrighted work itself. This reason (and others) is espoused by Adobe, Hugging Face, and StabilityAI in justifying how the training of models constitutes a "fair use" of copyrighted works under US law.[519] *(see section 5.3.1.B.1).*

3) Exclusion of copyrighted content from training datasets

[515] Katie Robertson, *OpenAI Strikes a Deal to License News Corp Content*, N.Y. TIMES (May 22, 2024), https://www.nytimes.com/2024/05/22/business/media/openai-news-corp-content-deal.html.
[516] *Id.*; Alexandra Bruell et al., *OpenAI, WSJ Owner News Corp Strike Content Deal Valued at over $250 Million*, WALL ST. J. (May 22, 2024), https://www.wsj.com/business/media/openai-news-corp-strike-deal-23f186ba.
[517] Brian Fung, *Thousands of authors demand payment from AI companies for use of copyrighted works*, CNN (July 20, 2023, https://www.cnn.com/2023/07/19/tech/authors-demand-payment-ai/index.html; Cheyenne DeVon, *Billie Eilish, Nicki Minaj, Jon Bon Jovi and over 200 artists call for protections against "predatory uses of AI,"* CNBC (Apr. 5, 2024), https://www.cnbc.com/2024/04/05/billie-eilish-nicki-minaj-200-artists-sign-letter-against-ai-music.html.
[518] Wes Davis, *AI companies have all kinds of arguments aginst paying for copyrighted content*, THE VERGE (Nov. 4, 2023), https://www.theverge.com/2023/11/4/23946353/generative-ai-copyright-training-data-openai-microsoft-google-meta-stabilityai.
[519] *Id.*; *see also* Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743 (2021) (machine learning use of data is transformative, not copying).

Some AI companies, such as Stability AI[520] and OpenAI,[521] have offered copyright holders the opportunity to opt out of having their work used in training datasets.[522] Artists have criticized this proposal because it requires owners of the copyrights to submit opt-out requests for each one of their copyrighted pieces, rather than opt in. OpenAI has also suggested that artists make use of "robots.txt," a decades-old method for website owners to indicate that they do not give permission for scraping data from their websites.[523] However, this is effective only when the artist also has control over the site hosting their images or material. The controversy has continued to grow. As of March 2024, one website that references AI-created content online estimates that around 32% of the 1,000 most popular websites use "robots.txt" to ban GPTbot,[524] up from only 7% in August 2023.[525] And in any case, this removal from datasets is forward-looking only; it will apply only to future training data and will not delete what models have already "learned" from past training data.

Pasquale and Sun have proposed a model that combines the two previous solutions: an opt-out for creators who do *not* want their works used and compensation for those who *do*, with the creation of a levy on AI providers for a fund to be administered by a central authority to compensate creators.[526] An option to opt out of training datasets or default into a compulsory licensing scheme is an intriguing possibility but less likely than either solution alone. That may leave a final possible resolution: Ensure that future generative AI models are trained only with data in the public domain. Various projects have been launched to gather datasets to train AI models wholly on public domain data, without using any copyrighted materials. For instance, Common Corpus is a public domain dataset released for training AI models.[527] The French nonprofit Fairly Trained has developed an LLM called KL3M that is believed to be the largest model trained on public data. Its training set is tiny compared to market leaders but may still offer a viable alternative.[528]

---

[520] Melissa Heikkila, *Artists can now opt out of the next version of Stable Diffusion*, MIT TECH. REV. (Dec. 16, 2022), https://www.technologyreview.com/2022/12/16/1065247/artists-can-now-opt-out-of-the-next-version-of-stable-diffusion//

[521] Kali Hays, *OpenAI offers a way for creators to opt out of AI training data. It's so onerous that one artist called it 'enraging'*, BUSINESS INSIDER (Sept. 29, 2023), https://www.businessinsider.com/openai-dalle-opt-out-process-artists-enraging-2023-9.

[522] This is distinct from users opting out of their data being harnessed for training purposes. Matt Burgess & Reese Rogers, *How to Stop Your Data from Being Used to Train AI*, WIRED (Apr. 10, 2024), https://www.wired.com/story/how-to-stop-your-data-from-being-used-to-train-ai/.

[523] OpenAI, *GPTBot*, OPENAI PLATFORM, https://platform.openai.com/docs/gptbot (last visited June 15, 2024).

[524] *Websites That Have Blocked OpenAI's GPTBot CCBot Anthropic Google Extended - 1000 Website Study*, ORIGINALITY.AI https://originality.ai/ai-bot-blocking (last visited Mar. 10, 2024).

[525] Hays, *supra* note 521; Kali Hays, *OpenAI's GPTBot and other AI web crawlers are being blocked by even more companies now*, BUSINESS INSIDER (Sept. 27, 2023), https://www.businessinsider.com/openai-gptbot-ccbot-more-companies-block-ai-web-crawlers-2023-9.

[526] Frank A. Pasquale & Haochen Sun, *Consent and Compensation: Resolving Generative AI's Copyright Crisis* Cornell Legal Studies Research Paper Forthcoming, (May 14, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4826695.

[527] Pierre-Carl Langlais, *Releasing Common Corpus: the largest public domain dataset for training LLMs*, HUGGING FACE (Mar. 20, 2024), https://huggingface.co/blog/Pclanglais/common-corpus.

[528] Kate Knibbs, *Here's Proof You Can Train an AI Model Without Slurping Copyrighted Content*, WIRED (Mar. 20, 2024), https://www.wired.com/story/proof-you-can-train-ai-without-slurping-copyrighted-content/.

3.3.2.B. Copyright-infringing output

Even though models generally create new outputs, it is possible that the content produced by a generative AI tool—such as an image, or even computer code—could turn out to be almost identical to that used in the training data. Given that generative AI models tend to memorize fragments of their training data, they might reproduce these fragments, potentially leading to charges of copyright infringement.[529] And in fact, with some wrangling, they might do more than reproduce fragments. Stanford University computer scientists were able to get a chatbot to regurgitate a full three and a half chapters from *Harry Potter and the Sorcerer's Stone*, as well as the entirety of a Dr. Seuss book.[530]

GitHub Copilot, an AI-powered coding assistant developed by GitHub (a subsidiary of Microsoft) and powered by the OpenAI Codex model, is being sued in US federal court by unnamed "J. Doe" programmers for violating the open-source software licenses under which code was published on Github.[531] GitHub Copilot was trained on Github's own public repositories of software code, much of which was published pursuant to software licenses that require anyone reusing the code to credit its creators. Some open-source developers alleged that GitHub Copilot's outputs reproduce copyrighted code without following the terms of these licenses.[532] This lawsuit has been allowed to proceed, meaning plaintiffs' theory of infringing outputs has been vindicated and has some merit.[533]

To mitigate the risk of copyright infringement for their customers, some generative AI providers have offered users indemnities. Google, Microsoft, Amazon, and OpenAI, among others, have pledged to indemnify certain users (particularly enterprise customers who do not fine tune or modify the model) for intellectual property claims they might face as a result of infringing outputs.[534] Although Microsoft and

---

[529] Ivo Emanuilov & Thomas Margoni, *Memorisation in generative models and EU copyright law: an interdisciplinary view*, KLUWER COPYRIGHT BLOG (Mar. 26, 2024), https://copyrightblog.kluweriplaw.com/2024/03/26/memorisation-in-generative-models-and-eu-copyright-law-an-interdisciplinary-view/.

[530] Peter Henderson et al., *Foundation Models and Fair Use*, arXiv (Mar. 28, 2023), https://arxiv.org/pdf/2303.15715. The authors conjectured that the ability of a model to regurgitate portions of long-form works was initially constrained by the size of a model's context window, which helps explain why they were able to get ChatGPT to regurgitate larger portions of *Harry Potter* text when using the GPT-4-based version of the chatbot (which has a larger context window). This also suggests that, as companies like Anthropic and OpenAI have updated their models with dramatically larger context windows, it may be possible to elicit even larger outputs of copyrighted works unless guardrails are strengthened. *See also* Chang et al., *supra* note 506.

[531] J.Doe 1 and J. Doe 2 v. Github, Inc., No. 3:22-cv-06823 (N.D. Cal. Nov. 3, 2022).

[532] *Id.* at 15–22; *see also* Ivo Emanuilov & Thomas Margoni, *Forget me not: memorisation in generative sequence models trained on open source licensed code* (Feb. 7, 2024), https://ssrn.com/abstract=4720990.

[533] J.Doe 1, et al. v. Github, Inc., No. 4:22-cv-06823-JST (N.D. Cal. Jan. 11, 2024).

[534] OpenAI Business terms (last updated Nov. 14, 2023), https://openai.com/policies/business-terms/ (Indemnification); Anthropic, PBC Commercial Terms of Service (effective Jan. 2024), https://www-cdn.anthropic.com/files/4zrzovbb/website/786ea99408c7b0c14684b6cf4e1b31d34b7a77aa.pdf (Indemnification); Brad Smith & Hossein Nowbar, *Microsoft announces new Copilot Copyright Commitment for customers*, MICROSOFT (Sept. 7, 2023), https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/; Peter Hallinan and Vasi Philomin, AWS MACHINE LEARNING BLOG (Nov. 29, 2023), https://aws.amazon.com/blogs/machine-learning/announcing-new-tools-and-capabilities-to-enable-responsible-ai-innovation/; Neal Suggs & Phil Venables, *Shared fate: Protecting*

Google have apparently extended this indemnification offer to all users, OpenAI and Anthropic have limited it to users of their premium or business tier.

3.3.2.C. Uncertain intellectual property status of AI-generated content

The question of who owns the intellectual property rights associated with the output of an AI model remains unresolved in most legal systems. For now, it could be considered that the individual writing the prompt owns the resulting output—provided that there is sufficient human contribution. Some leading providers, like OpenAI[535] and Anthropic,[536] affirm this view in their 2024 consumer terms of service—though with the proviso that this assignment of ownership is only "to the extent permitted by applicable law." This question may therefore be resolved differently in each jurisdiction or even each case, depending on local copyright laws, the originality of the particular output, and the extent of human involvement in generating the output. These issues are addressed under regulatory initiatives (*see chapter 5*).

3.4. Environmental, economical, and societal challenges

Beyond the risks associated with AI technology and its applications, and the legal challenges arising from its development, it is crucial to consider other long-term issues posed by the deployment of increasingly advanced generative AI models. These risks to society, sometimes referred to as "systemic risks,"[537] encompass several key areas: the potential for excessive market concentration, the impacts on employment, environmental consequences, and broader risks to humanity.

3.4.1. Concentration of market power

The generative AI market is expanding rapidly. The global generative AI market size is expected to reach $67.18 billion in 2024 and climb to $967.65 billion by 2032.[538] The technological research and consulting

customers with generative AI indemnification, GOOGLE CLOUD (Oct. 12, 2023), https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification.

[535] OpenAI Terms of Use (last updated Jan. 31, 2024), https://openai.com/policies/terms-of-use ("Ownership of Content . . . . to the extent permitted by applicable law, you (a) retain your ownership rights in Input and (b) own the Output. We hereby assign to you all our right, title, and interest, if any, in and to Output.").

[536] Anthropic Consumer Terms of Service (effective June 13, 2024), https://www.anthropic.com/legal/consumer-terms (amending a 2023 version that authorized only users to use outputs to say instead "we assign to you all of our right, title, and interest—if any—in Outputs.").

[537] Bengio et al., *International Scientific Report supra* note 7 § 4.3.

[538] *Generative AI Market Size, Share & Industry Analysis, by Model (Generative Adversarial Networks or GANs and Transformer-based Models), by Industry vs. Application, and Regional Forecast, 2024–2034*, FORTUNE BUSINESS INSIGHTS (May 27, 2024), https://www.fortunebusinessinsights.com/generative-ai-market-107837.

giant Gartner forecasts that, by 2026, 75% of businesses will be using generative AI to create synthetic customer data.[539] In 2023, funding for generative AI experienced a dramatic surge, to reach $25.2 billion.[540] Major players in the generative AI sector, including OpenAI, Anthropic, Hugging Face, and Inflection, have reported substantial fundraising rounds.[541] In this context, the market tends to become concentrated in the hands of a few powerful players, leading to several negative consequences.

3.4.1.A. Trends toward market concentration

In the generative AI market, barriers to entry are very high. Developers need access to vast volumes of data, computational resources, technical expertise, and capital. Large technology companies with such access are able to exploit economies of scale, economies of scope, and feedback effects (learning effects from user-generated data).[542] All this gives them an overwhelming advantage over smaller companies, making competition increasingly challenging for these smaller entities.[543] In 2023, the training costs for state-of-the-art AI models reached unprecedented levels.[544] For instance, OpenAI's GPT-4 required an estimated $78 million worth of compute resources to train, while Google's Gemini Ultra incurred training costs of $191 million for compute.[545]

Against this backdrop, leading generative AI developers have decided to partner with the few US tech giants dominating the market, such as IBM, Microsoft, Google, Nvidia, Meta, Apple, and Amazon Web Services.[546] For instance, Microsoft has invested millions of dollars in OpenAI, forming a partnership where Microsoft exclusively provides cloud data storage and API services.[547] Under this agreement, Microsoft integrates OpenAI's models into both consumer and enterprise products, as well as in developing "new categories of digital experiences." Similarly, Microsoft has entered into a multiyear partnership with Mistral, a French AI startup,[548] and invested $16 million into Mistral. This deal will integrate Mistral's open and commercial language models into Microsoft's Azure AI platform, making Mistral the second company to offer a commercial language model on Azure after OpenAI. The collaboration will focus on developing and deploying next-generation large language models, similar to Microsoft's existing partnership with

[539] Lori Perri, *3 Bold and Actionable Predictions for the Future of GenAI*, GARTNER (Apr. 12, 2024), https://www.gartner.com/en/articles/3-bold-and-actionable-predictions-for-the-future-of-genai.
[540] Stanford AI Index Report 2024 *supra* note 3.
[541] *Id.*
[542] Competition & Markets Authority, *supra* note 124.
[543] Elena Ponte et al., *Generative AI Raises Competition Concerns*, FED. TRADE COMM'N (June 29, 2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns.
[544] Stanford AI Index Report 2024 *supra* note 3.
[545] *Id.*
[546] FORTUNE BUSINESS INSIGHTS, *supra* note 538.
[547] MICROSOFT CORPORATE BLOGS, *supra* note 71.
[548] John K. Waters, *Microsoft Partners with Startup Mistral AI to Advance Next-Gen LLMs*, CAMPUS TECHNOLOGY (Feb. 27, 2024), https://campustechnology.com/Articles/2024/02/27/Microsoft-Partners-with-Startup-Mistral-AI-to-Advance-Next-Gen-LLMs.aspx.

OpenAI. For its part, Google has established a partnership with Anthropic, through which Google supplies cloud services essential for Anthropic's training, scaling, and deployment of its AI systems.[549]

On the technical front, the ability of the most powerful models to be fine-tuned for a wide range of tasks may result in a market in which just a few high-performance models dominate. The most powerful models are capable of meeting the majority of users' needs, thereby reducing the demand for new and diverse models. Consequently, they may capture the majority of the market share, centralizing control within a small number of entities.[550] In light of this risk, some advocate for open-sourcing AI models as a means to democratize the market and stimulate competition against established players, who typically prefer closed-source models to retain control over intellectual property and features. Open-source AI could, indeed, enhance competition by enabling many downstream developers to build upon existing models. However, such efforts are constrained by the limited availability of compute resources.[551] And in practice, some established AI companies have adopted open AI strategies to reinforce their market dominance, using openness as a means to solidify their control.[552]

In its latest publication, the UK Competition Markets Authority[553] emphasizes the increasing dominance of a few incumbent technology firms across the foundation models supply chain. Its report provides a figure illustrating interconnected relationships, where GAMMAN (Google, Amazon, Microsoft, Meta, Apple, Nvidia) firms have invested in or partnered with AI developers or other GAMMAN firms.

---

[549] Anthropic, *Anthropic Partners with Google Cloud*, ANTHROPIC (Feb. 2, 2023), https://www.anthropic.com/news/anthropic-partners-with-google-cloud.

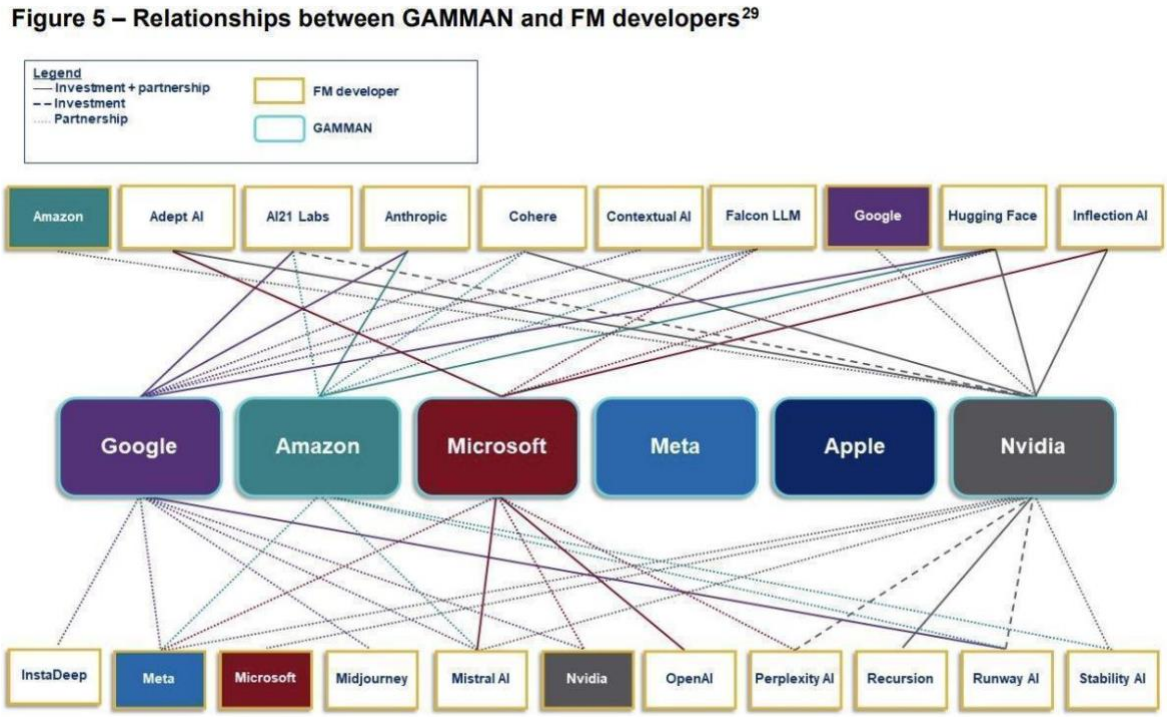
[550] Competition & Markets Authority, *supra* note 124.

[551] Seger et al. *supra*, note 192.

[552] David G. Widder et al., *Open (For Business): Big Tech, Concentrated Power, and the Political Economy of OpenAI*, SSRN (Aug. 18, 2023), https://ssrn.com/abstract=4543807.

[553] Competition & Markets Authority, *AI Foundation Models: Update Paper* (Apr. 11, 2024), https://assets.publishing.service.gov.uk/media/661941a6c1d297c6ad1dfeed/Update_Paper__1_.pdf. Since this update paper was released, Apple and OpenAI have announced a new partnership: Apple will install ChatGPT in its operating systems and Siri. S.M. Kelly, *The complicated partnership between Apple and OpenAI*, CNN (June 14, 2024), https://www.cnn.com/2024/06/14/tech/apple-openai-partnership.

FIGURE 10. Relationships between major tech companies and foundation model developers



source: Competition & Markets Authority, *AI Foundation Models: Update Paper* (Apr. 11, 2024), https://assets.publishing.service.gov.uk/media/661941a6c1d297c6ad1dfeed/Update_Paper__1_.pdf.[554]

In its report, the UK Competition and Markets Authority (CMA) lists the various existing partnerships in the AI industry, which encompass various domains, such as compute partnerships, data partnerships, and distribution partnerships.[555] Compute partnerships provide access to specialized supercomputing systems or chips, as seen in collaborations between Microsoft and OpenAI,[556] Amazon and Anthropic,[557] and Google and Anthropic.[558] Data partnerships involve one party gaining access to another's data resources, exemplified by Google's partnership with Reddit.[559] Distribution partnerships can take multiple forms. In some cases, a GAMMAN firm adds the partner's models to their library or provides access through their

[554] *Id.*

[555] *Id.* at 2.59.

[556] Microsoft Corporate Blogs, *supra* note 71.

[557] Amazon Staff, *What you need to know about the AWS AI chips powering Amazon's partnership with Anthropic*, AMAZON (Oct. 16, 2023), https://www.aboutamazon.com/news/aws/what-you-need-to-know-about-the-aws-ai-chips-powering-amazons-partnership-with-anthropic; Amazon Staff, *Amazon and Anthropic deepen their shared commitment to advancing generative AI*, AMAZON (Mar. 27, 2024), https://www.aboutamazon.com/news/company-news/amazon-anthropic-ai-investment.

[558] Anthropic, *supra* note 549; Google, *Announcing Anthropic's Claude 3 models on Google Cloud Vortex AI*, GOOGLE CLOUD (Mar. 4, 2024), https://cloud.google.com/blog/products/ai-machine-learning/announcing-anthropics-claude-3-models-in-google-cloud-vertex-ai.

[559] Rajan Patel, *An expanded partnership with Reddit*, THE KEYWORD (Feb. 22, 2024), https://blog.google/inside-google/company-announcements/expanded-reddit-partnership/.

developer tools, as evidenced by collaborations like Amazon and HuggingFace.[560] Other times, a GAMMAN firm integrates the partner's developer tools into its own platform or marketplace, such as Microsoft with Nvidia.[561] Additionally, GAMMAN firms may distribute a partner's AI infrastructure through their cloud marketplaces, as seen in the partnerships between Nvidia and Google,[562] and Nvidia and AWS.[563] These partnerships collectively enhance the development, deployment, and accessibility of advanced AI models, benefiting both developers and end users.

3.4.1.B. Negative effects of increased market concentration

The concentration of AI assets—encompassing data, hardware, and expertise—within a small group of global tech firms raises many concerns.[564] Such a situation may stifle healthy competition, impede innovation, and potentially result in elevated costs for accessing AI technologies. Firms with control over essential resources for developing AI models may restrict access to these resources to prevent competition. For instance, if, in the future, training AI models increasingly relies on proprietary data, smaller organizations lacking access to such data might encounter significant barriers to entry and growth.[565]

Partnerships between major players can strengthen their market power across the supply chain, further reducing competitive dynamics and solidifying their dominant positions. With a certain degree of control over the redistribution and utilization of their models, the most powerful players wield significant economic influence in determining access to their technology. They might use their market power to dictate their terms and technical standards to the rest of the market, affecting consumers and even influencing regulators. Potential negative impacts include reduced choices, lower quality, and higher prices. This asset concentration could also exacerbate economic and social inequality, as smaller enterprises and regions with limited access to AI resources may struggle to keep pace with or derive benefits from advancements in AI, resulting in a pronounced digital divide.

This report does not examine the legal implications of market concentration from the standpoint of competition law. However, it is noteworthy that competition authorities are carefully scrutinizing the

---

[560] Jeff Boudier et al., *Hugging Face and AWS partner to make AI more accessible*, HUGGING FACE (Feb. 21, 2023), https://huggingface.co/blog/aws-partnership.

[561] NVIDIA, *NVIDIA Teams with Microsoft to Build Massive Cloud AI Computer*, NVIDIA (Nov. 16, 2022), https://nvidianews.nvidia.com/news/nvidia-microsoft-accelerate-cloud-enterprise-ai; NVIDIA, *Microsoft and NVIDIA Announce Major Integrations to Accelerate Generative AI for Enterprises Everywhere*, NVIDIA (Mar. 18, 2022), https://nvidianews.nvidia.com/news/microsoft-nvidia-generative-ai-enterprises.

[562] NVIDIA, *Google Cloud and NVIDIA Expand Partnership to Scale AI Development*, NVIDIA (Mar. 18, 2024), https://nvidianews.nvidia.com/news/google-cloud-ai-development.

[563] NVIDIA, *AWS and NVIDIA Announce Strategic Collaboration to Offer New Supercomputing Infrastructure, Software and Services for Generative AI*, NVIDIA (Nov. 28, 2023), https://nvidianews.nvidia.com/news/aws-nvidia-strategic-collaboration-for-generative-ai; NVIDIA, *AWS and NVIDIA Extend Collaboration to Advance Generative AI Innovation*, NVIDIA (Mar. 18, 2024), https://nvidianews.nvidia.com/news/aws-nvidia-generative-ai-innovation.

[564] Philippe Lorenz & Kate Saslo, *Demystifying AI and AI Companies – What Foreign Policy Makers Need to Know About the Global AI Industry*, STIFTUNG NEUE VERANTWORTUNG (July 9, 2019), https://www.stiftung-nv.de/de/publikation/demystifying-ai-ai-companies-what-foreign-policy-makers-need-know-about-global-ai; Sanjay Chawla et al., *Ten Years after ImageNet: A 360° Perspective on Artificial Intelligence*, arXiv (Oct. 1, 2022), https://arxiv.org/pdf/2210.01797.

[565] Competition & Markets Authority, *supra* note 124 at 3.110.

behavior of AI companies. The EU Commission indicated that it is considering whether the partnership between Microsoft and OpenAI falls under the scope of its merger control powers.[566] In the United States, the Federal Trade Commission (FTC) issued orders to five companies (Alphabet, Amazon, Anthropic, Microsoft, and OpenAI) on January 25, 2024, requiring them to provide information regarding their recent investments and partnerships with generative AI firms and major cloud service providers.[567] The FTC is investigating whether these investments and partnerships by dominant companies could distort innovation and undermine fair competition—concerns that are similarly echoed by competition authorities in the EU. In April 2024, India's competition regulator launched a market study on AI and competition to "develop an in-depth understanding of the emerging competition dynamics in the development ecosystems of AI systems and implications of AI applications for competition, efficiency and innovation in key user industries."[568]

Finally, it is crucial to underscore that the concentration of the market in the hands of a few players presents not only a risk of restricting competition but also a significant geopolitical issue concerning technological world domination. The control and advancement of key technologies influences global economic dynamics and impacts national security, international relations, and the balance of power among nations. To date, the most advanced models have been primarily developed in the United States. According to the Stanford AI Index Report 2024, 61 notable AI models were developed by US-based institutions in 2023, significantly outnumbering the European Union's 21 and China's 15.[569] Other regions around the world are not witnessing similar levels of innovation. The concentration of advanced AI model development in a few developed countries raises concerns about potential dependence on these entities for critical technologies. This "Global AI Divide"[570] could escalate geopolitical tensions and diminish the autonomy of other nations. Lastly, from a purely technical perspective, the centralization of critical AI infrastructure in the hands of a few actors makes it a prime target for cyberattacks and espionage.

3.4.2. Impact on labor markets

The impact of generative AI on employment presents a significant challenge. While the deployment of AI across various professions offers numerous benefits—such as greatly enhancing efficiency by automating routine and repetitive tasks, and aiding in data analysis and decision-making processes—generative AI also

---

[566] European Commission, *Commission launches calls for contributions on competition in virtual worlds and generative AI*, EURO. COMM'N (Jan. 9, 2024), https://ec.europa.eu/commission/presscorner/detail/en/IP_24_85. On July 22, 2024, the EU Commission, U.K. Competition & Markets Authority, U.S. Department of Justice and Federal Trade Commission co-signed a *Joint Statement on Competition in Generative AI Foundation Models and AI Products*. *see:* https://competition-policy.ec.europa.eu/document/download/79948846-4605-4c3a-94a6-044e344acc33_en?filename=20240723_competition_in_generative_AI_joint_statement_COMP-CMA-DOJ-FTC.pdf

[567] *FTC Launches Inquiry into Generative AI Investments and Partnerships*, FED. TRADE COMM'N (Jan. 25, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships.

[568] *Competition Commission of India Launches Market Study on Artificial Intelligence and Competition*, COMPETITION COMM'N OF INDIA (Apr. 22, 2024), https://www.cci.gov.in/antitrust/press-release/details/385.

[569] Stanford AI Index Report 2024 *supra* note 3.

[570] Bengio et al. *International Scientific Report*, *supra* note 7 § 4.3.2.

has the potential to significantly disrupt labor markets. Experts examining this impact often conclude that, while generative AI may not lead to widespread job displacement, it will significantly alter the nature of many occupations.[571] Two primary concerns arise: the elimination of jobs due to automation and the exacerbation of economic inequalities.

3.4.2.A. Job loss and displacement

Currently, a significant share of workers (three in five) worry about losing their jobs entirely to AI in the next 10 years—particularly those who already work with AI.[572] Some studies conclude that AI tools (generative and non-generative) will create significant job losses.[573] The OECD has found that occupations at highest risk of being lost to automation from AI account for about 27% of employment.[574] Some studies have reached the conclusion that generative AI will affect at least 10% of the workloads of about 80% of the US workforce.[575] Meanwhile, other studies argue that, in the medium to long term, generative AI will create new jobs and industries and produce a net positive for jobs.[576] The critical question is whether new job creation will occur rapidly enough to offset the initial job losses.

Most experts agree that many jobs will *change*, as some aspects and components of jobs are complemented by AI.[577] For instance, an AI chatbot like ChatGPT reduces the *time and effort needed* for workers of all skill levels to complete tasks while it improves the quality of their output.[578] Similarly, coding assistants, like GitHub's Copilot, decrease by over 50% the time software developers need to complete a specific test task, with the most substantial gains seen among less experienced developers.[579] In customer service, agents using AI assistants saw a 14% increase in productivity, with novices and low-skilled workers benefiting the most.[580] Research also indicates that generative AI systems could make higher income jobs[581] and highly educated employees vulnerable to automation.[582] In this context, jobs involving routine and manual tasks

---

[571] Carl B. Frey & Michael Osborne, *Generative AI and the Future of Work: A Reappraisal*, 30 BROWN J. OF WORLD AFFAIRS 1–17 (2024), https://bjwa.brown.edu/30-1/generative-ai-and-the-future-of-work-a-reappraisal/.

[572] OECD, *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market,* OECD PUBLISHING, https://doi.org/10.1787/08785bba-en.

[573] *Id.*

[574] *Id.*

[575] *See* Tyna Eloundou et al., *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, arXiv (Aug. 21, 2023), https://arxiv.org/pdf/2303.10130; Edward W. Felten et al., *Occupational Heterogeneity in Exposure to Generative AI*, SSRN (Apr. 19, 2023), https://ssrn.com/abstract=4414065.

[576]Kweilin Ellingrud, *Generative AI and the Future of Work in America*, MCKINSEY GLOBAL INSTITUTE (July 26, 2023), https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america.

[577] OECD, *supra* note 572.

[578] Shakked Noy & Whitney Zhang, *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence*, 381 Science 6654 (July 14, 2023), https://www.science.org/doi/10.1126/science.adh2586.

[579] Sida Peng et al., *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*, arXiv (Feb. 13, 2023), https://arxiv.org/pdf/2302.06590.

[580] Erik Brynjolfsson et al., *Generative AI at Work*, NAT'L BUREAU OF ECONOMIC RESEARCH, Working Paper Series 31161 (Apr. 2023), http://www.nber.org/papers/w31161.

[581] Eloundou et al., *supra* note 575.

[582] Edward W. Felten et al., *How will Language Modelers like ChatGPT Affect Occupations and Industries?*, SSRN (Mar. 1, 2023), https://ssrn.com/abstract=4375268.

will probably become automated. Sectors such as telemarketing, administrative support, and technical support are especially vulnerable. But the continued integration of generative AI in everyday workplaces will also affect more sophisticated occupations, including legal services, investment, graphic design, and copywriting.[583] Professions based on writing and coding may face a greater risk of displacement compared to those grounded in scientific research or critical thinking.[584]

Still, it does not necessarily follow that because a job is vulnerable to automation it will immediately be automated. There are other factors for a business to consider before it decides to fully automate a task: technical feasibility and economic attractiveness.[585] Many businesses will find it too expensive to implement technology that will fully automate a job within a short time frame. They may have to wait years before it is reasonably affordable for them to do so. MIT's Computer Science & Artificial Intelligence Laboratory (CSAIL) studied the automation of vision-related tasks with today's current technology. It found that, at today's costs, "only 23% of worker wages being paid for vision tasks would be attractive to automate."[586] The study results do not preclude eventual labor displacement, but they stress that AI-caused labor change will be gradual and within a longer time frame.

Moreover, higher skilled jobs might just find their work augmented by the change, not automated. This is what could happen in the creative professions: Generative AI will reduce the difficulty of existing content-creating jobs.[587] ChatGPT has been shown to enhance the productivity of writers, particularly those with lower abilities.[588] And while generative AI can contribute to creative work, it is better suited to build upon existing ideas rather than generating entirely original narratives.[589] In this context, it is likely that generative AI tools will be valuable in assisting creators with their work, without replacing them.

Meanwhile, new professions are emerging in this developing industry, such as prompt engineer and prompt designer. Generative AI requires human input to prompt and select (and often edit) the desired output, with much of the actual creativity residing in this process.

Finally, in-person interactions remain valuable and cannot be easily replaced by machines.[590] While jobs not involving in-person communication may disappear, in-person communication is likely to become an increasingly important skill, for instance, in medical professions and in longstanding consumer relationships.

3.4.2.B. Rising inequalities

[583] OECD, *supra* note 572.
[584] Eloundou et al., *supra* note 575.
[585] Maja Svanberg et al., *Beyond AI Exposure: Which Tasks are Cost-Effective to Automate with Computer Vision?*, SSRN (Jan. 19, 2024), https://ssrn.com/abstract=4700751.
[586] *Id.*
[587] Frey & Osborne, *supra* note 571.
[588] Noy & Zhang, supra note 578.
[589] Frey & Osborne, *supra* note 571.
[590] *Id.*

In their book *The Second Machine Age*, Erik Brynjolfsson and Andrew McAfee show that information technology and computerization have significantly exacerbated income inequality through several mechanisms.[591] One key mechanism is "skill-biased technical change," which benefits more-skilled workers while replacing less-skilled workers.[592] This has led to widening income gaps between individuals with different educational backgrounds, such as those with a high school education versus college graduates. Additionally, there has been a shift from labor to capital, resulting in a decline in labor's share of income. Moreover, this trend has resulted in the rise of a relative few individuals who have leveraged digital technologies to reach massive audiences, propelling them into the top 0.1% of income earners.[593]

Indeed, studies show that generative AI can boost productivity, with varying effects on different groups of workers.[594] In an experiment conducted by Brynjolfsson et al.,[595] a large language model (LLM) was introduced in a call center to assist operators rather than replace them. The study found that less-skilled workers experienced the most significant benefits, with productivity increasing by about 35%. In contrast, the most-skilled workers saw almost no improvement. The LLM effectively captured and transferred the tacit knowledge of more experienced workers—such as problem-solving techniques and effective communication strategies—to the less-skilled workers.[596] Generative AI can help to narrow the productivity gap between the most experienced and less experienced workers. However, low-skilled workers might face job insecurity and wage stagnation, while high-skilled workers, especially in the tech sector, may benefit from increased demand and higher wages.

Ultimately, the true impact of generative AI on the job market will hinge on the decisions societies make regarding AI development. AI is more likely to displace workers when it is designed to replicate human skills and intelligence.[597] In such cases, there is a risk of concentrating wealth and power in the hands of a few individuals or organizations that control the capital. In addition, ordinary people, including those with significant expertise, may become less valued because machines would be performing their roles. This shift could lower wages, reduce the value of human work, and exacerbate economic inequality. Therefore, striving to imitate human capabilities may be misguided. Instead, societies should aim to increase wages and value by designing AI to complement human workers.[598] This approach could lead to substantial productivity gains without replacing human roles, thereby enhancing the value of human labor in the presence of AI.

---

[591] ERIK BRYNJOLFSSON & ANDREW MCAFEE, THE SECOND MACHINE AGE: WORK, PROGRESS, AND PROSPERITY IN A TIME OF BRILLIANT TECHNOLOGIES (1st ed.) (W. W. Norton & Company 2014).

[592] Erik Brynjolfsson, *The Jobs Equation*, THE ATLANTIC, https://www.theatlantic.com/sponsored/google-2023/the-jobs-equation-erik-brynjolfsson-qa/3872/ (last visited on June 16, 2024); *see also* Eli Berman et al., *Implications of Skill-Biased Technological Change: International Evidence*, NAT'L BUREAU OF ECONOMIC RESEARCH, Working Paper Series 6166 (Sept. 1997), https://www.nber.org/papers/w6166.

[593] *Id.*

[594] Brynjolfsson et al., *supra* note 580.

[595] *Id.*

[596] *Id.*

[597] Erik Brynjolfsson, *The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence*, arXiv (Jan. 11, 2022), https://arxiv.org/pdf/2201.04200.

[598] *Id.*

3.4.3. Environmental cost

Discussions on the environmental impact of generative AI can often be quite alarming. A study recently concluded that "a ChatGPT-like application" responding to an estimated 11 million requests per hour produces 12,800 metric tons of CO2 emissions each year.[599] Another study analyzing "the emissions of several AI systems (ChatGPT, BLOOM, DALL·E 2, Midjourney) relative to those of humans completing the same tasks" found that "an AI writing a page of text emits 130 to 1500 times less CO2e than a human doing so" and that "an AI creating an image emits 310 to 2900 times less."[600]

However, there is no widely accepted methodology for measuring the environmental impact of artificial intelligence and, more specifically, generative AI.[601] Certainly, it is possible to measure the energy consumed during training or inference and multiply that by the carbon intensity of the energy source used. However, it is difficult to have a precise idea of the impact of other aspects of the model life cycle, such as manufacturing hardware, heating and cooling data centers, or storing and transferring data. The environmental impact of AI may depend on factors that extend beyond the AI sector and even beyond the tech sector. For instance, BLOOM, a 176-billion parameter language model developed by Hugging Face, has a relatively low carbon footprint of around 25 metric tons of CO2 equivalent[602] because it was trained on a French supercomputer that primarily uses nuclear energy, which has a lower carbon footprint compared to fossil fuels.

Last, but not least, there remains a significant lack of data for a precise assessment of the full environmental costs of generative AI. Most AI developers do not report carbon emissions. The Stanford AI Index Report highlights that "most prominent model developers such as OpenAI, Google, Anthropic, and Mistral do not report emissions in training, although Meta does."[603] Therefore, estimates are primarily based on the limited information released by AI companies in their reports and data provided by local governments.

Within this framework, the following paragraphs aim to briefly shed light on two aspects of the environmental impact of generative AI: energy consumption and water consumption.

3.4.3.A. Energy consumption

---

[599] Andrew A. Chien et al., *Reducing the Carbon Impact of Generative AI Inference (today and in 2035)*, PROCEEDINGS OF THE 2ND WORKSHOP ON SUSTAINABLE COMPUTER SYSTEMS (July 9, 2023), https://doi.org/10.1145/3604930.3605705.

[600] Bill Tomlinson et al., *The Carbon Emissions of Writing and Illustrating Are Lower for AI than for Humans*, arXiv (Mar. 8, 2023), https://arxiv.org/pdf/2303.06219.

[601] For a comprehensive overview on the indirect effects of digitization, see Gauthier Roussilhe et al., *A long road ahead: a review of the state of knowledge of the environmental effects of digitization*, 62 CURRENT OPINION IN ENVIRONMENTAL SUSTAINABILITY 101269 (June 2023), https://doi-org.acces-distant.sciencespo.fr/10.1016/j.cosust.2023.101296.

[602] Alexandra Sasha Luccioni et al., *Estimating the Carbon Footprint of Bloom, a 176b Parameter Language Model* arXiv (Nov. 3, 2022), https://arxiv.org/pdf/2211.02001.

[603] Stanford AI Index Report 2024 *supra* note 3 at 156.

The energy consumption of generative AI can be considered in two distinct stages: the initial training of the model and the subsequent usage after deployment.

1) Training phase

The *training* stage of AI models, often recognized as the most energy-demanding phase, has attracted considerable attention in the field of AI sustainability research.[604] Training large AI models requires a substantial amount of computing power to handle vast datasets,[605] which translates into high energy consumption.[606] Hugging Face disclosed that its BLOOM model used 433 megawatt hours (MWh) of electricity for its training.[607] Comparatively, training GPT-3 consumed 1,287 MWh of electricity.[608] Since models need regular updates and retraining to incorporate the latest data and refine their functionality, more energy consumption is required. This persistent demand for updates, along with the associated energy consumption, intensifies the environmental impact of generative AI.

2) Inference phase

The energy consumption of the *inference* process—when an AI model generates a real-time response to a user's input—is usually seen as less substantial than during the training phase. However, the comparison of electricity consumption between the training and inference phases remains a subject of debate, as current research provides only limited insight into the comparative consumption of each phase.[609] And the elements communicated by AI companies can vary.

Data from Hugging Face shows that its BLOOM model uses substantially less energy for inference than for training.[610] This model uses 914 kilowatt hours (kWh) of electricity to process 230,768 requests, averaging out to 3.96 Watt hours (Wh) per request.[611] Similarly, Meta's Llama 65B model reportedly consumes between 2.8 and 5.5 Wh[612] per request, depending on the size of the batch.[613] For its part, Google

---

[604] Alex de Vries, *The growing energy footprint of artificial intelligence*, 7 JOULE 10, 2191–94 (2023), https://www.sciencedirect.com/science/article/abs/pii/S2542435123003653; David Patterson et al., *Carbon Emissions and Large Neural Network Training*, arXiv (Apr. 23, 2021), https://arxiv.org/pdf/2104.10350; Roberto Verdecchia et al., *A Systematic Review of Green AI*, arXiv (May 5, 2023), https://arxiv.org/pdf/2301.11047.

[605] Another study reported that the CO2 emissions created in training a single BERT model "is roughly equivalent to a trans-American flight." Emma Strubell et al., *Energy and Policy Considerations for Deep Learning in NLP*, arXiv (June 5, 2019), https://arxiv.org/pdf/1906.02243.pdf.

[606] Alexandra Sasha Luccioni et al., *Power Hungry Processing: Watts Driving the Cost of AI Deployment?*, arXiv (May 23, 2024), https://arxiv.org/pdf/2311.16863.

[607] Luccioni et al., *supra* note 602.

[608] de Vries, *supra* note 604; Patterson, *supra* note 604.

[609] de Vries, *supra* note 604.

[610] Luccioni et al., *supra* note 602.

[611] *Id.*

[612] Results in joules in the study: between 10^3 and 2 x 10^3 joules.

[613] Siddharth Samsi et al., *From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference*, arXiv (Oct. 5, 2023), https://arxiv.org/pdf/2310.03003.

has reported that 60% of its AI-related energy usage between 2019 and 2021 was due to inference.[614] In February 2023, Alphabet's chairman stated that engaging with a Large Language Model could "likely cost 10 times more than a standard keyword search."[615] Since Google estimated in 2009 that a typical keyword search required 0.3 Watt hour of energy,[616] it follows that each query made to a generative AI tool would consume 3 Watt hours. ChatGPT queries may consume around one gigawatt hour (GWh) each day, the equivalent of the daily energy consumption for about 33,000 US households.[617]

3.4.3.B. Water consumption

Data centers use water for cooling to prevent servers from overheating. The water consumption associated with AI training and inference processes can be substantial, impacting local water resources. In particular, training phases engage servers intensively and, therefore, create a need for additional cooling, which requires considerable water.[618]

One study suggests that GPT-3's training via Microsoft servers may have required 5.4 million liters of water, of which 700,000 liters were used directly on site for server cooling and 4.7 million liters for electricity consumption.[619] A lawsuit by local residents revealed that, in July 2022, the month before OpenAI completed training GPT-4, its Iowa-based data center cluster consumed approximately 6% of the district's water supply.[620] Additionally, as Google and Microsoft trained Bard and Bing models, they experienced significant increases in water usage, with annual spikes of 20% and 34%, respectively, according to the companies' environmental reports.[621]

3.4.3.C. Mitigation efforts

AI companies and researchers are pursuing advancements in energy-efficient model architectures, the adoption of renewable energy sources for data centers, and the implementation of carbon offset initiatives.

---

[614] David Patterson et al., *The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink*, arXiv (Apr. 11, 2022), https://arxiv.org/pdf/2204.05149.

[615] Jeffrey Dastin & Stephen Ellis, *Focus: For Tech Giants, AI Like Bing, Bard Poses Billion-Dollar Search Problem*, REUTERS (Feb. 22, 2023), https://www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22/.

[616] Google, *Powering a Google Search*, GOOGLE BLOG (Jan. 31, 2009), https://googleblog.blogspot.com/2009/01/powering-google-search.html; Dylan Patel & Afzal Ahmad, *The Inference Cost of Search Disruption – Large Language Model Cost Analysis*, SEMIANALYSIS (Feb. 9, 2023), https://www.semianalysis.com/p/the-inference-cost-of-search-disruption.

[617] Sarah McQuate, *How Much Energy Does ChatGPT Use?*, U. OF WASH. NEWS (July 27, 2023), https://www.washington.edu/news/2023/07/27/how-much-energy-does-chatgpt-use/.

[618] Pengfei Li et al., *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models*, arXiv (Oct. 29, 2023), https://arxiv.org/pdf/2304.03271.

[619] *Id.*

[620] Kate Crawford, *Generative AI's environmental costs are soaring — and mostly secret*, NATURE (Feb. 20, 2024), https://www.nature.com/articles/d41586-024-00478-x.

[621] *Id.*

Google has committed to running its data centers on carbon-free energy by 2030,[622] and Microsoft has pledged to become carbon negative by 2030.[623] The development of smaller, more efficient models can help reduce the energy consumption and carbon footprint of AI systems. Machine learning may help improve data centers' efficiency and reduce energy consumption.[624] For example, implementing DeepMind's machine-learning technology in Google data centers has successfully reduced energy usage for cooling by up to 40%.[625]

These initiatives may not be sufficient. At the World Economic Forum's annual meeting in Davos, (Switzerland) in January 2024, OpenAI's Sam Altman cautioned that the upcoming generation of generative AI systems will require significantly more power than anticipated, posing a challenge for existing energy infrastructure. Altman stated that anything short of a "breakthrough" in clean energy innovation may not offset the overwhelming energy costs of generative AI systems.[626] Ultimately, effective solutions can be developed only with full and transparent disclosure of the true environmental costs associated with training and operating AI models.

3.4.4. Artificial General Intelligence

Since the release of ChatGPT-3 in late 2022, there has been a growing focus on a possible existential threat associated with AI, commonly referred to as "x-risk." Many, especially within the AI community—including OpenAI's CEO Altman—are apprehensive about the potential threats of advanced versions of the technology,[627] especially a highly intelligent "rogue AI" that could surpass human oversight and potentially spin out of control in the future.[628]

---

[622] Google, *Net-zero Carbon*, GOOGLE SUSTAINABILITY, https://sustainability.google/operating-sustainably/net-zero-carbon/#:~:text=Run%20on%20carbon%2Dfree%20energy,where%20we%20operate%20by%202030&text=From%202010%20to%202022%2C%20we,than%2031%20million%20solar%20panels (last visited June 16, 2024).

[623] Microsoft, *Our Microsoft Sustainability Journey*, CORPORATE SOCIAL RESPONSIBILITY, https://www.microsoft.com/en-us/corporate-responsibility/sustainability-journey#:~:text=We're%20committed%20to%20being,we%20were%20founded%20in%201975 (last visited June 16, 2024).

[624] Matthew Smith et al., *Machine Learning-Based Energy-efficient Workload Management for Data Centers*, 2024 IEEE 21ST CONSUMER COMMUNICATIONS & NETWORKING CONFERENCE (CCNC) (Mar. 18, 2024), at 799–802, https://ieeexplore.ieee.org/document/10454842.

[625] Richard Evans & Jim Gao, *DeepMind AI Reduces Google Data Centre Cooling Bill by 40%*, GOOGLE DEEPMIND (July 20, 2016), https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/.

[626] Justine Calma, *Sam Altman Says the Future of AI Depends on Breakthroughs in Clean Energy*, THE VERGE (Jan. 19, 2024), https://www.theverge.com/2024/1/19/24044070/sam-altman-says-the-future-of-ai-depends-on-breakthroughs-in-clean-energy.

[627] Bender et al., *supra* note 221; Yuval Noah Harari, *Yuval Noah Harari argues that AI has hacked the operating system of human civilisation*, THE ECONOMIST (Apr. 28, 2023), https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation; Yuval Noah Harari, *Why Technology Favors Tyranny*, THE ATLANTIC (Oct. 2018), https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/.

[628] Bengio, *How Rogue AIs may Arise*, *supra* note 443.

The core of the so-called "existential risk" concern is the possibility that computers possessing intelligence that surpasses that of humans could lead to the destruction of most, if not all, human life. In an essay for *Financial Times*, Ian Hogarth, Chair of the UK AI Safety Institute,[629] made a case for AI companies to slow down the global race toward "God-like AI," which could be "a force beyond our control or understanding, and one that could usher in the obsolescence or destruction of the human race."[630] In a *New York Times* guest essay, Yuval Noah Harari, Tristan Harris, and Aza Raskin wrote, "We have summoned an alien intelligence. We don't know much about it, except that it is extremely powerful and offers us bedazzling gifts but could also hack the foundations of our civilization."[631]

In May 2023, the Center for AI Safety (CAIS), a nonprofit whose mission is to reduce societal risks from AI, released a statement which proclaimed: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[632] This statement has been endorsed by a diverse group of experts, including Altman and Anthropic's Dario Amodei, as well as renown computer scientists, nuclear physicists, legal scholars, economists, and philosophers. More recently, an open letter co-signed by several former OpenAI, Anthropic and Google DeepMind employees, titled "A Right to Warn about Advanced Artificial Intelligence,"[633] highlighted that the significant risks posed by advanced AI technology include exacerbating existing inequalities, enabling manipulation and misinformation, and losing control over autonomous AI systems, potentially leading to human extinction.

In contrast to these alarmist statements, others contest the assumption that artificial intelligence will pose an existential threat or feel ambivalent. While some assert that any potential existential risk necessitates immediate and significant measures to establish safeguards, others emphasize the importance of prioritizing the mitigation of existing, well-documented harms over hypothetical scenarios.[634] Additionally, some experts categorically dismiss the possibility of AI evolving into an existential threat. Given the highly contentious nature of this issue, only a few general observations will be made here.

3.4.4.A. Existential risk posed by Artificial General Intelligence

In a paper called "How Does Artificial Intelligence Pose an Existential Risk?" published in 2017, Karina Vold and Daniel Harris suggested that humans might create a super-intelligent machine that could outsmart all other intelligences, remain beyond human control, and potentially engage in actions that are contrary to

---

[629] *Introducing the AI Safety Institute*, Government of the United Kingdom - Department for Science, Innovation, and Technology, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

[630] Ian Hogarth, *We must slow down the race to God-like AI*, FINANCIAL TIMES (Apr. 12 2023), https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2.

[631] Yuval Harari et al., *You Can Have the Blue Pill or the Red Pill, and We're Out of Blue Pills*, N.Y. TIMES (Mar. 24, 2023), https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html.

[632] *Statement on AI Risk*, CENTER FOR HUMAN COMPATIBLE AI (SAFE.AI), https://www.safe.ai/work/statement-on-ai-risk/.

[633] Bengio et al., *A Right to Warn*, *supra* note 298.

[634] *See* NAIAC, *Statement On AI and Existential Risk*, AI.GOV (Oct. 2023), https://ai.gov/wp-content/uploads/2023/11/Statement_On-AI-and-Existential-Risk.pdf ("Arguments on existential risks should not detract from the necessity of addressing existing risks.").

human interests.[635] The prevailing narrative surrounding AI existential risk typically lies in the possibility of developing "Artificial General Intelligence" (AGI), or artificial super-intelligence (ASI).[636]

The answer to whether AGI represents a forthcoming development depends largely on how AGI is defined. Yet the definition of AGI remains subject to debate, and tech companies offer different definitions.[637] While OpenAI refers to "highly autonomous systems that outperform humans at most economically valuable work,"[638] IBM describes a situation where "artificial machine intelligence achieves human-level learning, perception and cognitive flexibility."[639] Researchers emphasize that, although there is no generally agreed upon definition of intelligence, "one aspect that is broadly accepted is that intelligence is not limited to a specific domain or task, but rather encompasses a broad range of cognitive skills and abilities."[640] Therefore, the term AGI can be used to refer to "systems that demonstrate broad capabilities of intelligence, including reasoning, planning, and the ability to learn from experience, and with these capabilities at or above human-level."[641] In fact, the term "artificial general intelligence" (AGI) gained popularity in the early 2000s to highlight the goal of advancing from "narrow AI"—which focuses on specific applications—to more comprehensive forms of intelligence.[642]

A recent study aggregated various definitions of AGI and produced a general definition that qualifies an AI system as AGI based on "generality" and "performance."[643] First, "generality" refers to the breadth of tasks a system can perform: A system is considered AGI if it can perform all or nearly all tasks.[644] Second, the "performance" criterion qualifies an AI system as AGI when its results are systematically better than those of a human *and* the system displays certain characteristics. Specifically, those characteristics include a reasoning structure identical to that of the human brain or thinking biases that incorporate mechanisms like those of a "consciousness."[645] Overall, AGI typically designates "an AI system that is at least as capable as a human at most tasks."[646]

---

[635] Karina Vold & Daniel R. Harris, *How does Artificial Intelligence Pose an Existential Risk? in* THE OXFORD HANDBOOK OF DIGITAL ETHICS (June 16, 2021), https://philpapers.org/archive/VOLHDA.pdf.

[636] Atoosa Kasirzadeh, *Two Types of AI Existential Risk: Decisive and Accumulative*, arXiv (Feb. 6, 2024), https://arxiv.org/pdf/2401.07836.

[637] Jeremy Baum & John Villasenor, *How close are we to AI that surpasses human intelligence?*, BROOKINGS (July 18 2023), https://www.brookings.edu/articles/how-close-are-we-to-ai-that-surpasses-human-intelligence/.

[638] OPENAI, *OpenAI Charter*, https://openai.com/charter/ (last visited June 16, 2024).

[639] Tim Mucci & Cole Stryker, *Getting ready for artificial general intelligence with examples*, IBM THINK 2024 (Apr. 18, 2024), https://www.ibm.com/blog/artificial-general-intelligence-examples/.

[640] Sébastien Bubeck et al, *Sparks of Artificial General Intelligence: Early experiments with GPT-4*, arXiv (Apr. 13, 2023), https://arxiv.org/pdf/2303.12712.

[641] *Id.*

[642] *Id.*

[643] Meredith Ringel Morris et al., *Levels of AGI for Operationalizing Progress on the Path to AGI*, arXiv (June 5, 2024), https://arxiv.org/pdf/2311.02462.

[644] There is some debate about whether physical tasks should be included or whether potential tasks should be limited to cognitive ones.

[645] Morris et al. *supra* note 643.

[646] *Id.*; OpenAI describes AGI as "highly autonomous systems that outperform humans at most economically valuable work." *See* OPENAI, *OpenAI Charter*, https://openai.com/charter (last visited June 16, 2024); *see also* Cade

3.4.4.B. Toward Artificial General Intelligence?

While some believe that AGI is a *conceivable* but *not certain* technological evolution,[647] others argue that AGI is already present in today's generative AI models.[648] This belief may partly rely on the observed "emergent capabilities" or "emergent behaviors" of AI models, even though these capabilities are still under discussion (*see section 3.2.5.B.*). Some AI companies, such as OpenAI[649] and Google DeepMind,[650] present AGI as an extension of the technologies they are currently developing. Ian Hogarth emphasized that "creating AGI is the explicit aim of the leading AI companies, and they are moving toward it far more swiftly than anyone expected."[651] Speaking at the World Economic Forum in Davos on January 18, 2024, Altman asserted that "the world is getting closer to AGI."[652] Yet OpenAI concedes on its website that "we cannot predict exactly what will happen and of course our current progress may hit a wall."[653]

Some researchers showed that AI models are already showing AGI "sparks."[654] Their results in certain fields (such as medicine or coding) are close to AGI in terms of outperforming humans. And the number of fields in which these models can perform is very broad: For instance, large language models cover anything involving text. The expectation is that with current techniques and more computing power, large models should be able to get closer to AGI—including the ability to plan and reason. One study even argues that such capabilities have a high probability of being achieved in less than a decade.[655]

3.4.4.C. Relativizing existential risk

Some experts, such as Yann LeCun and Andrew Ng, are skeptical about the near-term development of AGI due to the current limitations of AI technology, the complexity of replicating human intelligence, and the practical constraints of AI research.[656] To Yann LeCun, Meta's chief AI scientist, AGI represents an

---

Metz, *What's the Future for A.I.*, N.Y. TIMES (Mar. 31, 2023), https://www.nytimes.com/2023/03/31/technology/ai-chatbots-benefits-dangers.html.

[647] Morris et al. *supra* note 643.

[648] Blaise Agüera y Arcas & Peter Norvig, *Artificial General Intelligence is Already Here*, NOEMA (Oct. 2023), https://www.noemamag.com/artificial-general-intelligence-is-already-here/.

[649] OpenAI, *Planning for AGI and Beyond*, OPENAI BLOG (Feb. 24, 2023), https://openai.com/blog/planning-for-agi-and-beyond.

[650] Koray Kavukcuoglu, *Real-World Challenges for AGI*, GOOGLE DEEPMIND (Nov. 2, 2021), https://deepmind.google/discover/blog/real-world-challenges-for-agi/.

[651] Hogarth, *supra* note 630.

[652] World Economic Forum, *Technology in a Turbulent World with Sam Altman*, YOUTUBE (Feb. 2024), https://www.youtube.com/watch?v=JHPzQRTsb4A at 47:30.

[653] OpenAI, *Planning for AGI and beyond* (Feb. 2024), https://openai.com/index/planning-for-agi-and-beyond/.

[654] Bubeck et al., *supra* note 640.

[655] Yoshua Bengio et al., *Managing extreme AI risks amid rapid progress*, arXiv (May 22, 2024), https://arxiv.org/pdf/2310.17688.

[656]Yann LeCun & Andrew Ng, *Why the 6-month AI Pause is a Bad Idea*, YOUTUBE (Apr. 7, 2023), https://www.youtube.com/live/BY9KV8uCtj4.

"unattainable myth."[657] While LeCun acknowledges that machines may eventually surpass human intelligence, he believes this is a concern best relegated to the distant future and asserts that AI currently lacks the capacity to truly comprehend or make sense of the world. He even characterizes concerns about a potential threat to humanity "preposterously ridiculous"[658] and claims such concerns stem from human anthropomorphization of machines.[659] Aidan Gomez, CEO of the AI firm Cohere, adds that discussing the AI threat to human existence is "an absurd use of our time."[660]

Others have criticized the focus on the existential risk. For example, University of Washington Professor Emily has warned policymakers not to "fall for the distractions of AI hype."[661] Much of her criticism centers around the idea that risks of "rogue" AI are based on speculative fiction or "fantasies of techbros" in the far-off future, rather than academic research detailing harms of the present. Meredith Whittaker, president of the Signal Foundation and co-founder of the AI Now Institute, has also expressed skepticism about the existential risks associated with artificial intelligence. "There's no more evidence now than there was in 1950 that AI is going to pose these existential risks."[662] She argues that the focus on hypothetical existential threats from AI detracts from addressing the real, present-day harms caused by these technologies, especially the fact that they are controlled by a handful of corporations who ultimately make the decisions about them.[663] Others echo this thinking.[664]

In sum, the ability of advanced AI to perform a wide array of tasks and mimic human cognitive functions does not necessarily indicate the emergence of self-aware intelligence or the development of a goal that

---

[657] Yann LeCun, *I think the phrase AGI should be retired*, LinkedIn (2022), https://www.linkedin.com/posts/yann-lecun_i-think-the-phrase-agi-should-be-retired-activity-6889610518529613824-gl2F/.

[658] Melissa Heikkila, *Meta's AI leaders want you to know fears over AI existential risk are "ridiculous,"* MIT TECHNOLOGY REVIEW (June 20, 2023), https://www.technologyreview.com/2023/06/20/1075075/metas-ai-leaders-want-you-to-know-fears-over-ai-existential-risk-are-ridiculous/.

[659] Chris Vallance, M*eta's AI can 'recreate' clothes across different body shapes*, BBC (Feb. 9, 2022), https://www.bbc.com/news/technology-65886125.

[660] George Hammond, *The Future of AI: Ethical challenges*, FINANCIAL TIMES (June 16, 2023), https://www.ft.com/content/732fc372-67ea-4684-9ab7-6b6f3cdfd736.

[661] Emily M. Bender, *Policy makers: Please don't fall for the distractions of #AIhype*, MEDIUM (Mar. 29, 2023), https://medium.com/@emilymenonbender/policy-makers-please-dont-fall-for-the-distractions-of-aihype-e03fa80ddbf1.

[662] Will Douglas Heaven, *How existential risk became the biggest meme in AI*, MIT TECHNOLOGY REVIEW (June 19, 2023), https://www.technologyreview.com/2023/06/19/1075140/how-existential-risk-became-biggest-meme-in-ai/.

[663] Wilfred Chan, *Researcher Meredith Whittaker says AI's biggest risk isn't consciousness—it's the corporations that control them*, FAST COMPANY (May 5, 2023), https://www.fastcompany.com/90892235/researcher-meredith-whittaker-says-ais-biggest-risk-isnt-consciousness-its-the-corporations-that-control-them.

[664] Deb Raji said, "So much of the discussion was focused on concerns and promises outside the periphery of the most extreme dangers and benefits of AI rather than on adopting a clear-eyed understanding of the here and now. Speculation about the future of AI is fine as long as we don't spend all of our time daydreaming." Inioluwa Deborah Raji, *AI's Present Matters More than Its Imagined Future*, THE ATLANTIC (Oct. 4, 2023), https://www.theatlantic.com/technology/archive/2023/10/ai-chuck-schumer-forum-legislation/675540/.

opposes human interests.[665] Moreover, the likelihood of losing control over future advanced AI systems is a topic of significant debate, particularly given the current lack of extensive research assessing this risk.[666] There is, however, a consensus on the importance of establishing institutional knowledge and protocols to effectively address the rapidly advancing field of AI technology.

KEY TAKEAWAYS

- The actual and perceived risks associated with generative AI have garnered substantial attention and have been the focus of numerous studies. These risks span a broad spectrum and may manifest in the short, medium, or long term. Some originate from the inherent limitations of the technology in its current state, while others result from the ways humans choose to develop and use the technology. Other risks are linked to the legal, economic, labor, and environmental contexts. New risks will inevitably emerge with the advancement of future generative AI capabilities. Those include what are currently theoretical threats but ones that could pose significant long-term or even existential dangers to humanity.

- From a technical perspective, ensuring that an AI model is sufficiently robust is a complex challenge. AI models may exhibit unexpected behaviors or lack resilience against jailbreaking, where individuals manipulate the models to perform actions that violate usage restrictions. They may also be "misaligned," operating in ways that are inconsistent with the intended goals or values set by their creators or users, potentially causing harm. AI models can also unpredictably "hallucinate," presenting false information as factual, often with authoritative-sounding text and fabricated quotes and sources. AI generated outputs may contain biases or present a skewed view of reality due to their incompleteness or unrepresentative datasets. All these technical limitations are further exacerbated by the lack of transparency into the operations of generative AI models and their developers.

- From an ethical and social perspective, numerous additional risks arise from the potential applications enabled by this technology and the possibility that AI systems may be misused to cause harm. The ability of AI models to be employed for both intended and beneficial purposes, as well as unintended and harmful ones, is known as a "dual-use" risk. The advanced capabilities and widespread availability of generative AI models enable malicious actors to engage in harmful activities such as cybercrime, cyberattacks, or creating sexual deepfakes. The use of generative AI to create and widely disseminate disinformation is also a significant concern. The most alarming use cases include military applications and the potential for generative AI tools to be used in the creation of bioweapons. Finally, even in the absence of misuse, generative AI tools may exert excessive influence on the humans who interact with them and may lead to overreliance on these systems.

---

[665] *See* Fei-Fei Li & John Etchemendy, *AI LLM Is Not Sentient*, TIME (May 22, 2024), https://time.com/collection/time100-voices/6980134/ai-llm-not-sentient/.
[666] Bengio et al., *International Scientific Report, supra* note 7 at 53.

- Another concern arises from the rapid evolution of the technology, which may lead to the development of more capable AI agents that can autonomously interact with the world, plan ahead, and pursue goals. Such generative agents may exhibit emergent behaviors, meaning they can produce unexpected or surprising outputs. Some experts warn against the potential threats that advanced versions of the technology may pose in the future. They fear the emergence of a highly intelligent "rogue AI" that could surpass human oversight and potentially spin out of control. The core of the so-called "existential risk" concern is the hypothesis that computers possessing intelligence surpassing that of humans could lead to the destruction of most, if not all, human life. However, experts remain divided on the plausibility of the "loss of control" scenario.

- On the legal front, concerns result from the fact that developers train their models using extensive datasets often gathered through online web scraping, which may include personal data or copyrighted content. The issue is not only that personal data are used without the knowledge or consent of the individuals concerned, but also that generative AI models may memorize or leak personal data. Patterns or information structures identified within the data could enable malicious users to uncover personal details. Regarding copyrighted content, generative AI developers are frequently accused of violating copyright law by training AI models on copyrighted works without obtaining permission or compensating the copyright owners. Moreover, the content produced by a generative AI tool—such as an image or computer code—could sometimes be nearly identical to that used in the training data. And the question of who owns the intellectual property rights associated with the output of an AI model remains unresolved in most legal systems.

- More generally, risks to society, often referred to as "systemic risks," encompass several key areas: the potential for excessive market concentration, impacts on employment, and consequences for the environment. The generative AI market has very high barriers to entry, raising concerns that it may become concentrated in the hands of a few powerful players. Furthermore, generative AI has the potential to significantly disrupt labor markets, with the critical question being whether new job creation will occur rapidly enough to offset initial job losses. While most studies agree that many jobs will *change*, the true impact of generative AI on the job market is still debated and will depend on whether generative AI tools are designed to replicate human skills and intelligence or to complement human workers rather than replacing them. Additionally, discussions on the environmental impact of generative AI can often be alarming, given the energy and water needed to train AI models and manage data centers. However, there is not yet much data, and there is currently no widely accepted methodology for measuring the environmental impact of artificial intelligence, specifically generative AI.

- It is possible to question whether some generative AI models or systems should be considered more risky than others. This discussion centers on the most capable models and open-source models. Some experts and industry players view open-source models as potential sources of risk, due to the loss of control over the models, which can be exploited for malicious purposes. Conversely, advocates of open source see it as a solution to critical challenges, such as the excessive opacity of the most advanced models and market concentration around a few leading AI companies that have significantly invested in closed-source model development. As with other issues in this field, it

remains exceedingly difficult to resolve this debate and predict future outcomes. Furthemore, the term “frontier AI model” was coined to designate highly capable models that raise particular risks, based on the assumption that certain models with higher capacities inherently pose greater risks. However, the appropriate criteria for identifying such models is not clear-cut: For now, frontier models are identified by looking at the computational resources used for their training.

- Finally, the debate on the risks and challenges of generative AI is occurring within a context of significant uncertainty. Some risks, such as disinformation and environmental impact, are acknowledged, but measuring and estimating their consequences remains challenging. Other risks, such as the emerging capabilities of generative AI and potential “loss of control” scenarios, are debated but not universally accepted as proven. Additionally, anticipating the long-term consequences of the widespread deployment of generative AI is exceedingly difficult. Overall, the discussion of the risks and challenges of generative AI takes place within a paradoxical framework: AI companies themselves fuel the debate by publicly addressing these risks and publishing studies, yet these same companies remain relatively opaque and reluctant to disclose the information necessary for effectively evaluating the risks.

Chapter 4 Industry initiatives

The increasing public attention and evolving risks associated with generative AI have spurred AI companies to develop practices that mitigate risks while harnessing economic potential. It would be an overstatement to claim that individual measures by AI developers constitute industry-wide self-regulation, yet these initiatives may contribute to the creation of self-regulatory instruments. These emerging standards and practices are widely discussed and collaboratively refined within the AI community, often becoming recognized as best practices. Governments facilitate this process by encouraging transparency and collaboration among companies in disclosing their practices, developing and sharing technological advancements, and establishing unified standards. Such standards may subsequently be acknowledged by regulators, either as part of nonbinding frameworks, like the NIST framework (*see section 5.3.2.B.3.c.*), or within formal legal frameworks, such as the EU AI Act (*see section 5.1.2.*).

This chapter begins by offering a general overview of the practices commonly adopted by companies developing generative AI models and systems to address current risks and challenges (section 4.1). It then explores the collective initiatives within the industry that resemble self-regulation (section 4.2). It is important to emphasize that this chapter does not intend to provide a comprehensive technical analysis of industry practices. It neither cites all practices nor encompasses all AI companies. Nor is it intended to provide an overview of the current solutions developed by academic researchers to assess and mitigate the possible risks. Instead, the objective of this chapter is to highlight a few notable examples that are worth consideration by regulators and industry groups in their efforts to develop standards and best practices.

4.1. Industry practices

In June 2023, shortly after the European Parliament passed its own proposal for an AI Act (*see Appendix III*),[667] Stanford University's Center for Research on Foundation Models (CRFM) published an analysis grading compliance by foundation model providers with the draft regulation.[668] The CRFM listed the major companies already aligning with the provisions adopted by the European Parliament. This document was particularly insightful, offering a concise overview of AI companies' practices and highlighting those that demonstrated a responsible, virtuous attitude, and genuine transparency. It also illustrated that European drafters, specifically the EU Parliament, likely considered these industry practices when drafting the AI Act, highlighting a reciprocal influence between regulatory development and industry behavior.

[667] Amendments on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, EUR. PARL. DOC. (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

[668] Rishi Bommasani et al., *Do Foundation Model Providers Comply with the Draft EU AI Act?*, STAN. U. HUMAN-CENTERED ARTIFICIAL INTELLIGENCE, https://crfm.stanford.edu/2023/06/15/eu-ai-act.html (last visited June 16, 2024).

Leading AI developers have for years used public documents to describe core organizational commitments and establish internal standards for the development of AI systems. Examples include Google's AI Principles,[669] Microsoft's Responsible AI Standard,[670] and Meta's Five Pillars of Responsible AI,[671] which contain commitments like "avoid creating or reinforcing unfair bias" and internal standards like "Review defined Restricted Uses to determine whether the system meets the definition of any Restricted Use." In addition to these overarching documents, developers also publish numerous documents outlining a wide and rapidly evolving set of safety policies and practices relevant to specific models and applications.[672] Moreover, AI companies publish research that contributes to the development of new safety practices.[673]

Safety practices can be applied throughout the AI development-to-release pipeline, as well as post-release. While some practices are applicable across multiple stages, others are specific to certain phases. This section is structured into three groups of practices, organized approximately according to the generative AI development lifecycle. It is important to note, however, that this presentation has limitations. The actual practices and approaches adopted by AI companies can vary significantly, and the development lifecycle is not strictly linear. Therefore, while this presentation aims to provide a general framework for understanding, it does not fully capture the diversity and complexity of AI development and industry practices.

The groups of safety practices examined in this section are:

1. Pre-deployment: safety practices primarily applied prior to and during the training of a model,
2. Deployment: safety practices primarily applied after a model has been trained but before it is released, and
3. Post-deployment: safety practices primarily applicable after a model has been released.

4.1.1. Pre-deployment safety practices

The pre-deployment phase of the generative AI life cycle is a complex process. This phase can, roughly and schematically, be subdivided into two subphases, each of which presents opportunities for the assessment and mitigation of safety practices. The first is data preparation, which involves collecting and preparing data on which to train the AI model. The second is model development and training, the designing of the model's architecture and the training of the model using the prepared data. This discussion focuses on these two critical aspects of safety practices: data curation and model evaluation.

---

[669] Google, *Our Principles*, GOOGLE AI, https://ai.google/responsibility/principles/ (last visited June 16, 2024).

[670] Microsoft, *Principles and Approach*, MICROSOFT AI, https://www.microsoft.com/en-us/ai/principles-and-approach (last visited June 16, 2024).

[671] Meta, *Responsible AI: Driven by Our Belief that AI Should Benefit Everyone*, META AI, https://ai.meta.com/responsible-ai/ (last visited June 16, 2024).

[672] The self-reported risks include generating harmful content, hallucinations, disinformation that can be used in influence operations, material that undermines privacy and cybersecurity, and material that "reinforce[s] and reproduce[s] specific biases and worldviews, including harmful stereotypical and demeaning associations for certain marginalized groups." *GPT-4 Technical Report supra* note 289.

[673] Among others, Anthropic regularly publishes its own research on various AI safety-related topics: Anthropic, *Make Safe AI Systems, Deploy Them Reliably*, ANTHROPIC, https://www.anthropic.com/research (last visited May 19, 2024).

4.1.1.A. Data curation

Generative AI models derive their core capabilities from the data they are trained on, making the composition of that data a crucial determinant of the models' behavior and potential. Data governance—the policies, processes, and standards that developers use to ensure data availability, usability, integrity, and appropriateness—is, therefore, a matter of paramount importance.[674] Good data governance practices are important throughout the generative AI development lifecycle. Opportunities to mitigate risks associated with data arise during the pre-training and post-training stages of the pre-deployment phase and extend through to the post-deployment handling of user-provided data.[675] However, data governance is particularly crucial during the pre-training stage of the pre-deployment phase due to the foundational impact that pre-training has on the model's capabilities and the high costs associated with repeating it. Therefore, this section will concentrate on the curation of pre-training datasets.

A key activity of data governance for pre-training is data curation, or the process of ensuring the quality and appropriateness of training data.[676] It involves decisions regarding which sources of data should be included or excluded in the data aggregation process (source selection and data retention), the post facto removal of certain data from an aggregated dataset (data filtering), and the creation or augmentation of data to address gaps, imbalances, or other limitations (data augmentation/data synthesis).

Currently, the AI industry remains relatively opaque about its data governance practices.[677] However, it is possible to gain insights into key data governance practices employed in the industry by examining the available descriptions provided by companies and by drawing upon the broader literature on generative AI.[678]

1) Source selection

---

[674] Yacine Jernite et al., *Data Governance in the Age of Large-Scale Data-Driven Language Technology*. PROCEEDINGS OF THE 2022 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (June 20, 2022), https://doi.org/10.1145/3531146.3534637.

[675] Shayne Longpre et al., *Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them? An MIT Exploration of Generative AI* (Mar. 2024), MIT https://mit-genai.pubpub.org/pub/uk7op8zs/release/2; Shayne Longpre et al., *The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI*, arXiv (Nov. 4, 2023), https://arxiv.org/pdf/2310.16787.

[676] Jernite et al., *supra* note 675.

[677] Stanford's Center for Research on Foundation Models (CRFM) noted in its inaugural October 2023 *Foundation Model Transparency Index* that "[d]evelopers are least transparent with respect to the resources required to build foundation models." Rishi Bommasani et al., *The Foundation Model Transparency Index*, arXiv (Oct. 19, 2023), https://arxiv.org/pdf/2310.12941. In its May 2024 update to the *Index*, CRFM noted that, with some minor improvement, this trend had continued, noting that "[d]ata remains a key area of opacity and "[d]evelopers display a fundamental lack of transparency with respect to data." The *Index*'s author argues that "[t]hese low scores reflect the ongoing crisis in data provenance, wherein companies share no information about the license status of their datasets, preventing downstream developers from ensuring they are complying with such licenses." Rishi Bommasani et al., *The Foundation Model Transparency Index v1.1*, STAN. U. (May 2024), https://crfm.stanford.edu/fmti/paper.pdf.

[678] Comprehensive data on this matter have previously been aggregated by the CRFM in its *Foundation Model Transparency Index*. *See id.*

In the context of ongoing discussions about the biases inherent in machine-learning models (*see section 3.2.3.*), AI experts have emphasized the importance of careful data source selection.[679] For instance, scholars have raised concerns about the representativeness of internet datasets, such as Common Crawl,[680] which are widely used in training generative AI models (*see section 2.2.2.A.*). Some argue that, despite the vast size of these datasets, they may not adequately capture the diversity of ideas, perspectives, and experiences found across different communities, particularly those underrepresented online.[681] One path for addressing this limitation is to deliberately include additional sources that better reflect underrepresented communities on the internet. Developers can also opt to exclude data sources that are already well-represented by other data sources, ensuring a more balanced and diverse dataset.

Currently, it is unclear how and to what extent careful source selection is effectively implemented in the collection of datasets that are used to train many leading generative AI systems. For example, OpenAI, in its model training FAQ, simply notes that it uses "data from different places including public sources, licensed third-party data, and information created by human reviewers."[682] It mentions that its DALL·E 2 image generation model was trained on "hundreds of millions of captioned images from the internet," without identifying specific search or selection protocols for those images.[683] Anthropic and Google are not specific about the sources they have used to collect training data.[684] Meta doesn't provide much details about the sources of training data for Llama 3 (*see Figure 11*).[685]

FIGURE 11. Examples of source selection practices

Company	Model Family	Relevant Statements on Source Selection	Documents Reviewed
Anthropic	Claude 3	• "*Claude 3 models are trained on **a proprietary mix of publicly available information** on the Internet as of August 2023, as well as **non-public data from third parties, data provided by data-labeling***	*The Claude 3 Model Family: Opus, Sonnet, Haiku*

[679] Bender et al. *supra* note 221; Emily M. Bender & Alexander Koller, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*, *Annual Meeting of the Association for Computational Linguistics*, PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (2020) at 5185–98; Angelina McMillan-Major et al., *Documenting Geographically and Contextually Diverse Data Sources: The BigScience Catalogue of Language Data and Resources*, arXiv (Jan. 25, 2022), https://arxiv.org/pdf/2201.10066.

[680] Stefan Baak, *Training Data for the Price of a Sandwich*, MOZILLA INSIGHTS (Feb. 6, 2024), https://foundation.mozilla.org/fr/research/library/generative-ai-training-data/common-crawl/.

[681] Bender et al. *supra* note 221. Bad sourcing practices can have much worse consequences in extreme cases. An investigation showed that images of child sexual abuse material were present in LAION-5B. See Schuhmann et al., *supra* note 25; Thiel, *supra* note 25.

[682] OpenAI, *Enterprise privacy at OpenAI*, OPENAI, https://openai.com/enterprise-privacy (last visited June 16, 2024).

[683] *GPT4 Technical Report*, *supra* note 289.

[684] Bommasani et al., *The Foundation Model Transparency Index*, *supra* note 678.

[685] Meta, *Introducing Meta Llama 3: The most capable openly available LLM to date*, META BLOG (Apr. 18, 2024), https://ai.meta.com/blog/meta-llama-3/. About Llama 2, *see*: Hugo Touvron et al., *Llama 2: Open Foundation and Fine-Tuned Chat Models*, arXiv (July 18, 2023), https://arxiv.org/pdf/2307.09288.

		services and paid contractors, and data we generate internally*….* ● *"When Anthropic obtains data by crawling public web pages,* ***we follow industry practices with respect to robots.txt instructions and other signals*** *that website operators use to indicate whether they permit crawling of the content on their sites. In accordance with our policies, Anthropic's crawler does not access password-protected or sign-in pages or bypass CAPTCHA controls, and we conduct diligence on the data that we use.* ● *"Anthropic operates its crawling system transparently, which means* ***website operators can easily identify Anthropic visits*** *and signal their preferences to Anthropic."*[686]	
OpenAI	GPT-4	● "*GPT-4 is a Transformer-style model pre-trained to predict the next token in a document,* ***using both publicly available data (such as internet data) and data licensed from third-party providers****…* ● *"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains* ***no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar****."*[687] ● "*GPT-4 has learned from* ***a variety of licensed, created, and publicly available data sources, which may include publicly available personal information****."*[688]	*GPT-4 Technical Report*
Meta	Llama 3	● "*Llama 3 was pretrained on over 15 trillion tokens of data* ***from publicly available sources****.* ● *"The fine-tuning data includes publicly available instruction datasets, as well as over 10M human-annotated examples. Neither the pretraining nor the fine-tuning datasets include Meta user data…. The pretraining data has a cutoff of March 2023 for the 8B and December 2023 for the 70B models respectively."*[689] ● "*To train the best language model, the curation of a large, high-quality training dataset is paramount. In line with our design principles, we invested heavily in pretraining data. Llama 3 is pretrained on over 15T tokens* ***that were all***	*Llama 3 Model Card, Introducing Meta Llama 3: The most capable openly available LLM to date (blog post)*

[686] Anthropic, *The Claude 3 Model Family: Opus, Sonnet, Haiku*, ANTHROPIC (Mar. 2024), https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf [hereinafter *The Claude 3 Model Family*].

[687] *GPT4 Technical Report*, *supra* note 289.

[688] *Id*.

[689] Aston Zhang, *LLama 3 Model Details*, GITHUB (Apr. 20, 2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

		collected from publicly available sources. *Our training dataset is seven times larger than that used for Llama 2, and it includes four times more code. To prepare for upcoming multilingual use cases, over 5% of the Llama 3 pretraining dataset consists of high-quality non-English data that covers over 30 languages.*"[690]	
Google	Gemini	• "*Gemini models are trained on a dataset that is both multimodal and multilingual. Our pre-training dataset uses* ***data from web documents, books, and code****, and includes image, audio, and video data*."[691] • "*Our pre-training dataset includes* ***data sourced across many different domains, including web documents and code****, and incorporates image, audio, and video content. For the instruction tuning phase, we fine tuned Gemini 1.5 models on a collection of multimodal data (containing paired instructions and appropriate responses), with further tuning based on human preference data*."[692]	*Gemini: A Family of Highly Capable Multimodal Models, Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*
Technology Innovation Institute	Falcon	• "*We assembled a pretraining dataset of 3,500 billion tokens,* ***predominantly sourced from our work on RefinedWeb*** *(Penedo et al., 2023)–a massive filtered and 'deduplicated' web dataset*."[693] • "*We train small 1B models on 30B tokens, with the pretraining data split between web data and a specific curated category. We sample training on 1, 10, 25, 50, 75, and 100% of the targeted category. We only consider a one-dimensional approach, and mix web data with a single category of curated data. We split our categories in books, conversations, and technical data as outlined in Table 5.* • *"For the individual corpora making these categories,* ***we draw inspiration from The Pile*** *(Gao et al., 2020)* ***which we enhance with data from Reddit*** *(Baumgartner et al., 2020) for the conversational category. Our web data is taken from RefinedWeb (Penedo et al., 2023) and we process curated sources through a similar pipeline, applying filtering and deduplication to*	*The Falcon Series of Open Language Models*

[690] Meta, *Introducing Meta Llama 3*, *see supra* note 684.
[691] Gemini Team, Google, *Gemini: A Family of Highly Capable Multimodal Models*, https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf (last visited June 16, 2024).
[692] Gemini Team, Google, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf, (last visited June 16, 2024).
[693] Ebtesam Almazrouei et al., *The Falcon Series of Open Language Models*, arXiv (Nov. 29, 2023), https://arxiv.org/pdf/2311.16867.

		make for a fair comparison."[694]	

2) Data filtering

Data filtering is a powerful tool for guiding model behavior. Instead of excluding data from a source as a whole, filtering allows for a subset of harmful or otherwise undesirable data collected from the source to be removed. It is not necessarily desirable to remove all harmful content from a training dataset, as this can limit a model's ability to detect harmful data or cause it to amplify undesirable biases. But managing the relative composition of ideas and perspectives in a dataset is widely understood as an important technique for mitigating bias and errors. Poor data quality not only raises significant safety concerns, such as fairness and bias issues, but also severely impacts model performance.

Given the vast scale of data used to train generative AI models, filtering relies primarily on a combination of heuristic rules and algorithms to identify and remove unwanted data. For example, OpenAI provides the following general description of how its developers make use of classifier models to find and filter out violent and sexual images from the training data used to create the DALL·E 2 model:

"First, we create a specification for the image categories we would like to label; second, we gather a few hundred positive and negative examples for each category; third, we use an active learning procedure to gather more data and improve the precision/recall trade-off; and finally, we run the resulting classifier on the entire dataset with a conservative classification threshold to favor recall over precision. To set these thresholds, we prioritized filtering out all of the bad data over leaving in all of the good data."[695]

Practices vary and evolve rapidly. Although Google appears to have removed personal data from its training data for PaLM 2, it does not describe the removal of violent or other kinds of harmful data.[696] For its Gemini family of models, Google states that it applied "quality filters to all datasets, using both heuristic rules and model-based classifiers," in addition to performing "safety filtering to remove harmful content based on our policies."[697] Regarding pre-training datasets for its Gemini 1.5 model family specifically, Google says that it applied "safety filtering to our pre-training data for our strictest policies."[698]

In training its Llama 2 model, Meta justifies its decision not to filter out such content, saying it is to avoid "the potential for the accidental demographic erasure sometimes caused by over-scrubbing."[699] OpenAI acknowledged similar risks to explain its decision to use a lower threshold for certain "broad filters for

[694] *Id.*
[695] OpenAI, *DALL-E 2 Pre-training Mitigations*, https://openai.com/index/dall-e-2-pre-training-mitigations/ (last visited June 16, 2024).
[696] Rohan Anil et al., *PaLM 2 Technical Report*, arXiv (Sept. 13, 2023), https://arxiv.org/pdf/2305.10403.
[697] Gemini Team, Google, *Gemini: A Family of Highly Capable Models*, arXiv (June 17. 2024), https://arxiv.org/pdf/2312.11805.
[698] Gemini Team, Google, *Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context*, https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf (last visited June 16, 2024).
[699] Touvron et al., *supra* note 685.

sexual and violent imagery" for its DALL·E 3 model than the threshold it used for DALL·E 2. This change addressed a previous bias against generating images of women introduced by applying overly broad filters, according to OpenAI. The broader filters were replaced by "more specific filters on particularly important sub-categories" of the same content type.[700] For its GPT-4 model, OpenAI reports that it "reduced the prevalence of certain kinds of content that violate our usage policies (such as inappropriate erotic content) in our pre-training dataset," in addition to "removing personal information from the training dataset where feasible."[701]

3) Data augmentation

Data augmentation has emerged as a central technique for improving the performance of image models.[702] By creating many copies of images that have been rotated, cropped, recolored, or otherwise augmented, developers can dramatically increase the size and comprehensiveness of image training data.[703] The quality and size of training datasets for *text* models can also be increased through these kinds of relatively simple augmentations, like "shuffling sentences, changing the positions of words, replacing words with close synonyms, inserting random words, and deleting random words."[704] Data augmentation can be especially helpful in mitigating privacy-related harms, with notable potential for applications in healthcare, where patient privacy is central.[705]

4) Data synthesis

While data augmentation, or "partial" data synthesis, refers to approaches that retain obvious features of real-world data, "full" data synthesis involves the creation of data that is entirely distinct from the original.[706] Generating *synthetic* data[707] that mimics real-world data has become a common practice,[708] with some predicting the majority of data used to train AI systems will be synthetic within only a few years.[709]

---

[700] OpenAI, *DALL-E 3 System Card*, (Oct. 3, 2023), https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf.

[701] *GPT-4 Technical Report, supra* note 289.

[702] Amazon Web Services, *What Is Data Augmentation?*, https://aws.amazon.com/what-is/data-augmentation/#:~:text=Text%20data%20augmentation (last visited July 22, 2024).

[703] Connor Shorten & Taghi M. Khoshgoftaar, *A Survey on Image Data Augmentation for Deep Learning*, 6 J. BIG DATA 1, 1-48 (2019).

[704] Amazon Web Services *supra* note 702.

[705] Mauro Giuffrè & Dennis L. Shung, *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy* (October 9, 2023) NPJ DIGITAL MEDICINE, 6(1):186, https://pubmed.ncbi.nlm.nih.gov/37813960/.

[706] Jigyasa Grover & Rishabh Misra, *Keeping it Low-Key: Modern-Day Approaches to Privacy-Preserving Machine Learning*, *in* DATA PROTECTION IN A POST-PANDEMIC SOCIETY (Chaminda Hewage et al. eds., 2023), https://doi.org/10.1007/978-3-031-34006-2_2; Peter Lee, *Synthetic Data and the Future of AI*, 110 CORNELL L. REV. (Feb. 10, 2024), https://ssrn.com/abstract=4722162.

[707] Lee, *supra* note 706.

[708] Cade Metz & Stuart A. Thompson, *What to Know About Tech Companies Using A.I. to Teach Their Own A.I.*, N.Y. TIMES (Apr. 6, 2024), https://www.nytimes.com/2024/04/06/technology/ai-data-tech-companies.html.

[709] Neil Savage, *Synthetic Data Could Be Better Than Real Data*, NATURE (Apr. 27, 2023), https://www.nature.com/articles/d41586-023-01445-8; *The Claude 3 Model Family supra* note 686; *see also* Rob Toews, *Synthetic Data Is About to Transform Artificial Intelligence*, FORBES (June 12, 2022),

It offers an affordable and scalable alternative for obtaining large datasets for training machine-learning models.[710] The utility of synthetic data is often a function of how well it resembles the attributes and statistical properties[711] of real-world data. The measure of this similarity is the data's *fidelity*. Ideally, synthetic data is entirely free of any potentially sensitive information found in real-world data while still retaining enough fidelity for it to be useful for a given use case.[712]

Among other advantages, synthetic data provides a way to address biases found in real-world data that reflect social inequities or that result from incomplete or unbalanced collection. By synthesizing more diverse, representative, or inclusive data, developers can address biases or fill in gaps in real-world datasets, resulting in more fair and useful models.[713] However, synthetic data can also have the opposite impact. Synthetic data that too closely mimics the statistical properties of real-world data can reproduce real-world biases in the models it is used to train.[714] And because synthetic data is often generated by AI models that closely resemble the models they are used to train, failure to properly remove bias from synthetic data creates a risk of feedback loops that perpetuate or even magnify biases.[715]

In general, overreliance on synthetic data can risk incorporating too much fabricated content that skews the model's behavior in undesirable ways.[716] Furthermore, there is always a risk that the original real-world data can be reconstructed from synthetic data. This becomes problematic if the original data is sensitive, such as medical records or financial transactions.[717] It is generally best practice to limit the amount of synthetic content within training datasets to prevent the model from becoming too detached from reality.[718]

---

https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=3e78acab7523.

[710] ODSC, *How Synthetic Data Can Be Used for Large Language Models* (Sept. 18, 2023), https://opendatascience.com/how-synthetic-data-can-be-used-for-large-language-models/#:~:text=Synthetic%20data%20is%20artificial%20data,fewer%20legal%20issues%20and%20costs.

[711] Fernando Lucini, *The Real Deal About Synthetic Data*, MIT SLOAN MANAGEMENT REVIEW, Winter 2022, (Oct. 20, 2021), https://sloanreview.mit.edu/article/the-real-deal-about-synthetic-data/.

[712] Grover & Misra, *supra* note 706.

[713] Sam Forsdick, *Artificial advantage: can synthetic data make AI less biased?* (Aug. 1, 2022), https://www.raconteur.net/technology/artificial-advantage-can-synthetic-data-make-ai-less-biased/.

[714] Grover & Misra, *supra* note 706; Lee, *supra* note 696; Leinar Ramos & Jitendra Subramanyam, *Maverick Research: Forget About Your Real Data — Synthetic Data Is the Future of AI*, GARTNER RESEARCH (June 24, 2021), https://www.gartner.com/en/documents/4002912.

[715] Lucini, *supra* note 711.

[716] Mostly AI, *What Is Synthetic Data?*, https://mostly.ai/synthetic-data/what-is-synthetic-data#:~:text=Synthetic%20data%20is%20generated%20by,create%20statistically%20identical%2C%20synthetic%20data (last visited June 15, 2024).

[717] "For text data, a basic tradeoff exists between fidelity and privacy: as the synthetic data is made increasingly similar to the real-world data on which it is based, the risk correspondingly increases that the original real-world data can be reconstructed from the synthetic data. If that original real-world data is sensitive—medical records or financial transactions, say—this is a problem. A core challenge for synthetic text data, therefore, is not just to maximize fidelity in a vacuum, but rather to maximize fidelity while preserving privacy." Rob Toews, *Synthetic Data Is About to Transform Artificial Intelligence*, FORBES (June 12, 2022), https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=93c55a375238.

[718] Kim Bozzella, *The Pros and Cons of Using Synthetic Data for Training AI*, FORBES (Nov. 20, 2023), https://www.forbes.com/sites/forbestechcouncil/2023/11/20/the-pros-and-cons-of-using-synthetic-data-for-training-ai/?sh=743520de10cd.

4.1.1.B. Model evaluation and testing

Once the model is pre-trained, possible risks must be identified and understood. Model evaluation techniques include benchmarking and red teaming.

1) Benchmarking

Benchmarking is the process of measuring and comparing the performance of different models using standardized datasets, metrics, and tasks.[719] The goal is to objectively assess, in a reproducible manner, the capabilities and limitations of various models. Benchmarking involves using widely accepted datasets, well-defined evaluation metrics, and specific tasks to quantify the quality of generated outputs. However, benchmarking is not without issues. Experts criticize many current benchmarks as "static," too narrowly focused on individual capabilities, or out of step with the real ways generative AI models are used today.[720]

The evaluation of generative AI models also suffers from a lack of standardization. The benchmarking efforts by leading generative AI companies often employ widely varying approaches and criteria.[721] The 2024 Stanford AI Index notes that "leading developers, including OpenAI, Google, and Anthropic, primarily test their models against different responsible AI benchmarks."[722] This practice "complicates efforts to systematically compare the risks and limitations of top AI models."

Sophisticated benchmarking methodologies must still be developed to accurately identify risks arising from generative AI's increasingly complex behaviors.[723] New efforts are emerging, with notable work being done by MLCommons,[724] a consortium of more than 125 global AI engineering members and affiliates from across industry, academia, and the nonprofit sector, which aims to develop more comprehensive and dynamic benchmarking standards (*see section 4.2.4.*).

2) Red teaming

"Red teaming," or the process of engaging adversarially with systems in order to expose their limitations and vulnerabilities, has emerged as a leading approach to evaluating the risks of generative AI models and systems.[725] The practice of red teaming originated as wargame exercises during the Cold War and was later widely adopted by cybersecurity practitioners.[726] By simulating "attacks" on systems, red teaming can

---

[719] Bengio et al., *International Scientific Report supra* note 7 at 35.

[720] Kyle Wiggers, *Here's Why Most AI Benchmarks Tell Us So Little*, TECHCRUNCH (Mar. 7, 2024), https://techcrunch.com/2024/03/07/heres-why-most-ai-benchmarks-tell-us-so-little.

[721] Timothy R. McIntosh et al., *Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence*, arXiv (Feb. 15, 2024), https://arxiv.org/pdf/2402.09880.

[722] Stanford AI Index Report 2024, *supra* note 3.

[723] McIntosh, et al., *supra* note 721.

[724] MLCommons, *Better AI for Everyone*, https://mlcommons.org/ (last visited June 16, 2024).

[725] Bengio et al., *International Scientific Report supra* note 7 at 36.

[726] Jessica Ji, *What Does AI Red-Teaming Actually Mean?,* GEO. U. CENTER FOR SECURITY AND EMERGING TECHNOLOGY (Oct. 24, 2023),

uncover weaknesses in existing safety measures or elicit previously unanticipated and undesirable behaviors. The term is now also used to describe a foundational technique in generative AI model safety, with leading providers such as OpenAI,[727] Anthropic,[728] or Inflection AI ,[729] emphasizing its importance in the development and evaluation processes.

While the industry has promoted red teaming as a crucial tool for addressing risks in generative AI systems, the lack of transparency around internal practices means there are limited concrete details on how red teaming is implemented and its actual effectiveness.[730] Efforts to pierce industry opacity around red teaming activities have found that there is considerable variation in both goals and processes, as well as a lack of standards or systematic procedures for disclosing or taking action on findings.[731] Regulators[732] and

---

https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/#:~:text=As%20the%20name%20of%20the,red%2Dteaming%E2%80%9D%20from%20cybersecurity.

[727] OpenAI describes red teaming as "an integral part of [its] iterative deployment process."OPENAI, *Red Teaming Network*, https://openai.com/blog/red-teaming-network (last visited June 16, 2024).

[728] In his July 2023 testimony to the U.S. Senate Judiciary Committee on AI oversight, Anthropic CEO Dario Amodei highlighted the importance of red teaming in his company's efforts to mitigate risk, calling it "essential, and particularly important right now." *Written Testimony of Dario Amodei, Ph.D., Co-Founder and CEO, Anthropic, For a hearing on "Oversight of A.I.: Principles of Regulation" Before the Judiciary Comm., Subcomm. on Privacy, Technology, and the Law, United States Senate*, 118th Cong. (July 25, 2023), https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf.

[729] Inflection describes red teaming as "the engine at the heart of our evaluation framework" and "the best indication of how a model will perform in real-world situations." INFLECTION AI PRESS, https://inflection.ai/press (last visited June 16, 2024); Inflection AI, *Frontier Safety*, https://inflection.ai/frontier-safety#area7 (last visited June 16, 2024).

[730] In addition to the general lack of transparency noted elsewhere in this report and best documented in the *Foundation Model Transparency Index* from the Stanford CRFM, there is a documented lack of transparency around the specific issue of red teaming. Dr. Natasha Bajema, Senior Research Associate at the James Martin Center for Nonproliferation Studies at the Middlebury Institute of International Studies, argues that tech developers engaged in red-teaming efforts "tend to hire contractors and require them to sign nondisclosure agreements," conducting red-teaming exercises "behind closed doors" and providing the public with little detail of their results. Natasha Bajema, *Why Are Large AI Models Being Red Teamed?*, IEEE SPECTRUM (Mar. 15, 2024), https://spectrum.ieee.org/red-team-ai-llms.

[731] In surveying real-world cases of AI red-teaming exercises and the existing literature on red teaming and its application to generative AI, Carnegie Mellon University researchers found a "lack of consensus around the scope, structure, and assessment criteria for AI red-teaming," as well as "the resulting decisions the activity instigates (e.g., reporting, disclosure, and mitigation)." Michael Feffer et al., *Red-Teaming for Generative AI: Silver Bullet or Security Theater?*, arXiv (May 15, 2024), https://arxiv.org/abs/2401.15897.

[732] The Biden Administration's October 30, 2023, Executive Order (E.O.) 14110 on *Safe, Secure, and Trustworthy Deployment and Use of Artificial intelligence*, instructing the U.S. National Institute of Standards and Technology (NIST) to develop "rigorous standards for extensive red-team testing to ensure safety before public release." *Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*, WHITE HOUSE (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/; In December 2023, NIST issued a public request for information in support of its response to the E.O., with responses due by February 2, 2024. At the time of writing, NIST's final guidelines, due July 26, 2024, are still forthcoming. *NIST Calls for Information to Support Safe, Secure and Trustworthy Development and Use of AI Technologies*, NAT'L INST. OF STANDARDS & TECH. (Dec. 19, 2023), https://www.nist.gov/news-events/news/2023/12/nist-calls-information-support-safe-secure-and-trustworthy-development-and.

academia[733] are currently working to advance red teaming for generative AI from a promising risk mitigation approach to a fully developed and mature risk mitigation framework. The following paragraphs highlight some key features of red teaming in its current state.

a) A technique for identifying risks and vulnerabilities

Red teaming involves the use of input from humans or other AI models to identify and understand AI system vulnerabilities and risks. Generative AI "red teams" can involve a company's internal personnel, contracted third parties, or uncontracted public participants. These red team members test various inputs to see if they can prompt the AI system being tested to produce harmful outputs or exhibit undesirable behaviors. These tests are highly diverse, ranging from attempts to generate harassing or violent content to trying to leverage AI coding assistants to automate breaches in cybersecurity mechanisms.[734]

Generative AI red teaming can be performed at multiple stages of generative system development. Red teaming at the base model level (e.g., GPT-4) helps developers understand a model's fundamental capabilities and limitations to inform downstream application development, while red teaming at the application level (e.g., ChatGPT) helps identify risks related to the specific context within which a model is being used.[735]

Google, Anthropic, Hugging Face, OpenAI, and Meta all describe implementing red-teaming practices to varying degrees to test the robustness and safety of their AI models. Google announced that it had "conducted novel research on safety risks and developed red-teaming techniques to test for a range of potential harms" on its Gemini 1.5 Pro model.[736]At Anthropic, a "Trust & Safety team" conducted a comprehensive multimodal red-team exercise to evaluate Claude 3 and ensure alignment with the company's Acceptable Use Policy.[737] Hugging Face has a robustness research team for red teaming BLOOM, but external red teaming is not explicitly implemented.[738] OpenAI carries out red-teaming campaigns with external experts in a wide variety of fields.[739] Meta implements red teaming both internally and externally, conducting a series of red-teaming exercises with over 350 people, including internal employees, contract workers, and external vendors.[740]

---

[733] Within academia, a notable effort was made by researchers at Carnegie Mellon University. To help the field "move toward a more robust toolbox of evaluations for generative AI," the researchers published a detailed table of questions and considerations for AI red teaming. The table, offered as a starting point for "careful co-design, development, and evaluation," can be found in the appendix of this report. M. Feffer, *supra* note 731.

[734] Weidinger et al., *supra* note 252 at 214–29.

[735] Ram Shankar Siva Kumar, *Microsoft AI Red Team: Building the Future of Safer AI*, MICROSOFT (Aug. 7, 2023), https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/#:~:text=Guidance%20and%20resources%20for%20red%20teaming.

[736] Demis Hassabis & Sundar Pichai, *Our next-generation model: Gemini 1.5*, GOOGLE BLOG (Feb. 15, 2024), https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#ethics-safety.

[737] *The Claude 3 Model Family*, *see supra* note 686.

[738] Kaijie Zhu et al., *PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts*, arXiv (July 16, 2024), https://arxiv.org/pdf/2306.04528.

[739] *OpenAI Red Teaming Network*, OPENAI BLOG (Sept. 19, 2023), https://openai.com/blog/red-teaming-network.

[740] Touvron et al., *supra* note 685.

b) Requirements for an effective red-teaming strategy

Red teaming's effectiveness is dependent on a red team's ability to anticipate and imitate a wide variety of potential user behaviors. The scale of this challenge is significant. The number of users of leading generative AI models has grown dramatically, with some exceeding 100 million monthly active users. This creates a dramatic asymmetry between the number of users with the potential to create harmful or wanted behaviors and the relatively small number of red teamers working to prevent those behaviors.[741] The ways users will interact with generative AI tools is inherently hard to predict—a problem that continues to grow as more models accept multimodal inputs.[742] Red teaming generative AI models is further complicated by the probabilistic nature of the technology. Compared to traditional cybersecurity red teaming, generative AI red teaming must contend with "multiple layers of non-determinism" that can mean a single input can result in different outputs.[743]

While it may be impossible to develop a test set covering all possible ways that users might elicit undesirable or harmful behavior from a model, red teaming seeks to address at least a meaningful subset of a model's risk profile.[744] To do this, red teamers must anticipate novel behavior by users across a wide spectrum of attack types, which requires not only technical competence but also considerable imagination. As such, a red team's effectiveness can be significantly affected by its composition. For instance, OpenAI acknowledged that the composition of the red-team testing its GPT-4 model included "50 experts from domains such as long-term AI alignment risks, cybersecurity, biorisk, and international security," but the members were primarily Western, English-speaking, and drawn from specialized educational and professional backgrounds. As a result, the process likely suffered from bias that led it to anticipate certain risk types over others.[745]

If there is an insufficient diversity of expertise, perspectives, and imaginative outlooks, red teams could miss important risks. Anthropic indicates that, in the future, thorough red teaming will require "at least many

---

[741] Anna Tong, *ChatGPT Traffic Slips for Third Month in a Row*, REUTERS (Sept. 7, 2023), https://www.reuters.com/technology/chatgpt-traffic-slips-again-third-month-row-2023-09-07/.

[742] For example, the size of the context window for Google's Gemini mode grew from 32,000 tokens for the 1.0 version to 1 million tokens for the 1.5 version. Demis Hassabis & Sundar Pichai, *Our next-generation model: Gemini 1.5*, GOOGLE BLOG (Feb. 15, 2024), https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#context-window.

[743] Ram Shankar Siva Kumar, *Announcing Microsoft's Open Automation Framework to Red Team Generative AI Systems*, MICROSOFT SECURITY BLOG (Feb. 22, 2024), https://www.microsoft.com/en-us/security/blog/2024/02/22/announcing-microsofts-open-automation-framework-to-red-team-generative-ai-systems/#:~:text=2.%20Generative%20AI%20is%20more%20probabilistic%20than%20traditional%20red%20teaming.

[744] Christian Schlarmann & Matthias Hein, *On the Adversarial Robustness of Multi-Modal Foundation Models*, arXiv (Aug. 21, 2023), https://arxiv.org/pdf/2308.10741.pdf; Anna Tong, *ChatGPT Traffic Slips for Third Month in a Row*, REUTERS (Sept. 7, 2023), https://www.reuters.com/technology/chatgpt-traffic-slips-again-third-month-row-2023-09-07/.

[745] By OpenAI's own account, this "likely influenced both how red teamers interpreted particular risks as well as how they probed politics, values, and the default behavior of the model" and "privilege[d] the kinds of risks that are top of mind in academic communities and at AI firms." OPENAI, *supra* note 345 at 5.

dozens of hours of deliberate red teaming per topic area, by world class experts specifically focused on these threats (rather than students or people with general expertise in a broad domain)."[746] For companies deploying highly capable generative AI systems, effective red teaming may require the hiring or contracting of many specialized teams. Larger and more established companies are better positioned to meet these financial and organizational requirements. However, given the scale and complexity of the challenge, even highly resourced teams at major companies are likely to encounter gaps in expertise in certain areas.[747]

One response to this limitation has been to scale up the size of the red teams through outsourcing and crowdsourcing. OpenAI launched its Red Teaming Network in September 2023.[748] The company describes the move as part of a broader transition "from a focus on internal adversarial testing at OpenAI, to working with a cohort of external experts."[749] Similarly, in October 2023, Inflection AI said it was commissioning outside experts.[750] That same month, Microsoft launched its AI "Bug Bounty" program, offering "bounties" or rewards of up to $15,000 for those who identify vulnerabilities in its "AI-powered Bing experience."[751]

c) Automated Red Teaming (ART)

Another emerging approach to solve the scaling and imagination demands inherent in red teaming AI models is "automated red teaming" (ART).[752] Increased automation within the red-teaming process could allow for the automated generation of adversarial attacks at scale by significantly decreasing the amount of required human resources and increasing the efficiency, frequency, and scope of red-teaming activities.[753] In addition to offering a way to more thoroughly test the safety of generative systems, ART could also reduce the emotional burden on human red teams, whose work can expose them to toxic and harmful content.[754]

---

[746] *Anthropic's Responsible Scaling Policy, Version 1.0.,* ANTHROPIC (Sept. 19, 2023), https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf.

[747] *Why Red Teams Play a Central Role in Helping Organizations Secure AI Systems*, GOOGLE (July 2023), https://services.google.com/fh/files/blogs/google_ai_red_team_digital_final.pdf.

[748] *OpenAI Red Teaming Network*, OPENAI BLOG (Sept. 19, 2023), https://openai.com/blog/red-teaming-network.

[749] *Id.*

[750] *Our Policy on Frontier Safety*, INFLECTION AI (Oct. 30,2023), https://inflection.ai/frontier-safety.

[751] *Microsoft AI Bounty Program*, MICROSOFT (Apr. 11, 2024), https://www.microsoft.com/en-us/msrc/bounty-ai.

[752] Perez et al., *supra* note 261.

[753] As one example, Anthropic cites the "substantial time (i.e., 100+ hours)" that subject matter and LLM experts must spend to execute a well-defined red-teaming plan" as motivation for developing "automate evaluations based on expert knowledge, and the tooling to run those evaluations to make them repeatable and scalable." *Frontier Threats Red Teaming for AI Safety*, ANTHROPIC (July 26, 2023), https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.

[754] Bhaktipriya Radharapu, et al., *AART: AI-Assisted Red-Teaming with Diverse Data Generation for New LLM-powered Applications*, CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (2023), https://arxiv.org/pdf/2311.08592.pdf.

Alongside the numerous ART methods being developed by academia,[755] industry has put forth frameworks that it hopes can capitalize on the advantages of automation. For instance, Google Research has developed the AI-assisted Red-Teaming (AART), a partial ART system that combines human review with an automated review to generate adversarial attacks using a set of guidelines, or "recipes."[756] Google says the approach has already been used to improve the safety of several released products.[757] For its part, Microsoft's AI Red Team released the Python Risk Identification Toolkit (PyRIT), an open automation framework to conduct red-team exercises on generative AI systems.[758]

To better evaluate the extent to which various ART methods improve upon human teams, researchers affiliated with the Center for AI Safety, Microsoft, and several leading academic institutions have developed the "HarmBench" benchmark. HarmBench is a standardized evaluation framework for automated red teaming. Researchers identified several desirable properties previously unaccounted for in red-teaming evaluations and systematically designed HarmBench to meet these criteria.[759]

FIGURE 12. An example of standardized evaluation framework for automated red teaming: HarmBench

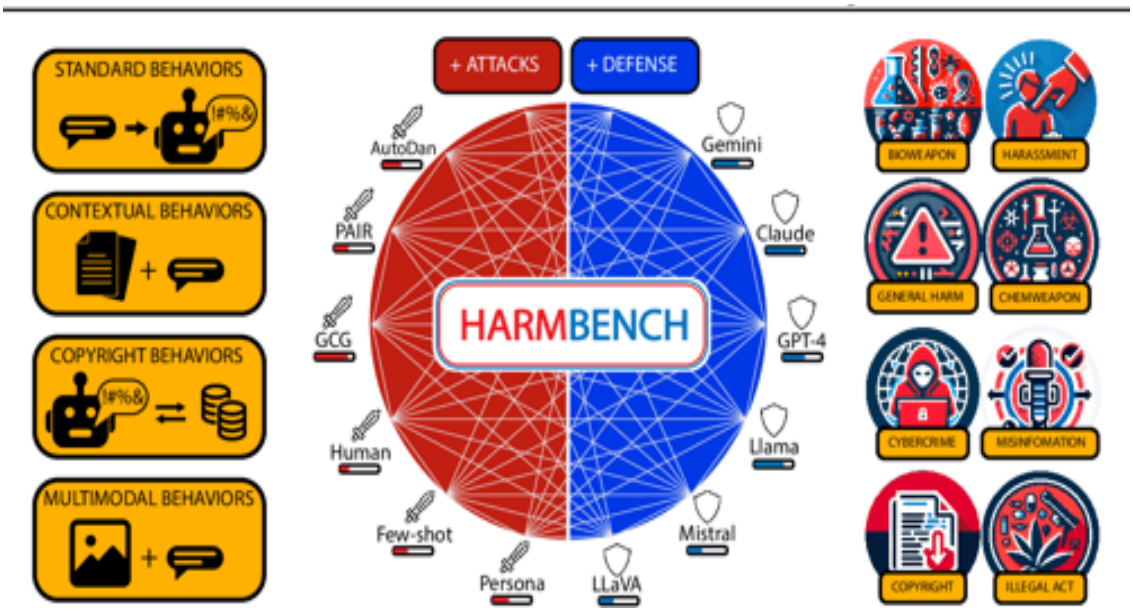
[755] The following represent a small and not necessarily representative sample of methods developed within academia. Gelei Deng et al., *MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots*, arXiv (Oct. 25, 2023), arXiv, https://arxiv.org/pdf/2307.08715; Alex Mei et al., *Assert: Automated Safety Scenario Red Teaming for Evaluating the Robustness of Large Language Models*, arXiv (Nov. 11, 2023), https://arxiv.org/pdf/2310.09624.

[756] Radharapu et al., *supra* note 754.

[757] Nicholas Carlini et al., *Are aligned neural networks adversarially aligned?*, arXiv (May 6, 2024), https://arxiv.org/pdf/2306.15447; Mantas Mazeika et al., *HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal*, arXiv (Feb. 27, 2024), https://arxiv.org/pdf/2402.04249.

[758] Ram Shankar Siva Kumar, *Announcing Microsoft's Open Automation Framework to Red Team Generative AI Systems*, MICROSOFT SECURITY BLOG (Feb. 22, 2024), https://www.microsoft.com/en-us/security/blog/2024/02/22/announcing-microsofts-open-automation-framework-to-red-team-generative-ai-systems/.

[759] Mazeika, et al., *supra* note 757.



Source: Mantas Mazeika et al., *HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal*, arXiv (Feb. 27, 2024), https://arxiv.org/pdf/2402.04249.

However, ART has limitations. At least in the case of partially automated systems like Google's AART, human input in the form of subject matter expertise and review is still essential. It requires "developers [to] work with other stakeholders to define the dimensions of the adversarial evaluation, such as ways that attackers structure queries, regions where the application is to be deployed, categories of harm that are high-risk for the application, or expanding on previously identified weaknesses."[760] It is also difficult to assess the extent to which partial-ART and full-ART techniques are currently used across industry.

d) Limitations of red teaming

Red teaming, whether performed by humans, AI, or a mixture of the two, is only a tool for *identifying* risks. However, it is not a means for *mitigating* risks. Once identified by red teams, vulnerabilities and risks must be met with rigorous measurement work to understand their pervasiveness[761] and must be paired with technical and design interventions. Moreover, while red teaming is essential for anticipating risks, it also imposes a significant cost burden on developers.

[760] Radharapu et al., *supra* note 754.

[761] As Microsoft notes, "the role of RAI red teaming is to expose and raise understanding of risk surface and is not a replacement for systematic measurement and rigorous mitigation work. It is important that people do not interpret specific examples as a metric for the pervasiveness of that harm." *Planning Red Teaming for Large Language Models (LLMs) and Their Applications*, Microsoft Learn (Nov. 6, 2023), https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming.

Within this context, it is very difficult to determine the reliability of companies' claims regarding their red-teaming efforts, especially since there is no mandate requiring companies to consistently follow through on their commitments. For example, the release of Google's Gemini 1.0 model family raised questions about Google's red-teaming practices after the model generated images—such as racially diverse Nazi-era German soldiers—that were historically inaccurate.[762] Google's own documents suggest that its red-teaming efforts were not applied evenly, or at all, to some models within the Gemini family prior to launch.[763] The Gemini 1.0 Technical Report states that the company "carried out red teaming on a December 2023 Gemini API Ultra checkpoint" with no mention of similar red teaming on checkpoints of the other members of the Gemini model family. This is an example of how, at present, the utility of red-teaming as a mitigation practice is undercut by a lack of standardization and enforcement.

4.1.1.C. Model alignment

AI *alignment* is the process of ensuring that an AI model's behavior and outcomes are "aligned" with the goals and values established by its designers. Therefore, the primary objective of alignment is to enhance the performance and reliability of AI models from the developers' perspective. Addressing the risk of misalignment (*see section 3.1.1.B.*) involves adopting better practices in the development phase.[764]

For instance, the risk of inequities or biases in pre-trained models may be addressed by improving supervised fine-tuning practices (see *section 2.2.3.B.*).[765] For example, Aya, a multilingual open-source model developed by Cohere, was created by fine-tuning the 13-billion parameter mT5 pre-trained model using a highly-curated labeled dataset of 101 languages. The result was a model capable of interacting in less common ("low-resource") languages.[766]

Similarly, reinforcement learning is a widely used technique to align generative AI models with desired objectives (see *section 2.2.3.A.*). However, although widely used and effective, reinforcement learning with human feedback (RLHF) may present limitations due to its reliance on feedback provided by humans. Similar to content moderation on social media, it relies on a well-established industry of vendors who employ human annotators. After the launch of ChatGPT, media reported that OpenAI had used the San

[762] Alex Heath, *Google CEO says Gemini diversity errors are 'completely unacceptable,'* THE VERGE (Feb. 27, 2024), https://www.theverge.com/2024/2/28/24085445/google-ceo-gemini-ai-diversity-scandal-employee-memo.

[763] Gemini Team, *Gemini: A Family of Highly Capable Multimodal Models* (April 25, 2024), GOOGLE DEEPMIND, https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

[764] The prominence of these techniques is captured in Google's Gemini 1.0 Technical Report, which states that "Our modeling mitigation of safety risks, applied across Gemini Advanced and Gemini API Ultra models, is mostly through post-training [...], encompassing supervised fine-tuning (SFT) and reinforcement learning through human feedback (RLHF) using a reward model." Google states that its approach to mitigation for Gemini 1.5 Pro remained "mostly the same" as for Gemini 1.0. Gemini Team, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* , GOOGLE DEEPMIND, https://arxiv.org/pdf/2403.05530 (last visited June 20, 2024).

[765] Bergmann, *supra* note 151.

[766] Ahmet Üstün et al., *Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model*, arXiv, (Feb. 12, 2024), https://arxiv.org/abs/2402.07827.

Francisco-based vendor Sama to outsource labeling to workers in Kenya,[767] paying them less than $2 per hour for their work on a toxicity classifier model.[768] While the work Sama did for OpenAI was not directly related to reinforcement learning and predated the release of ChatGPT, the incident provided a rare window into how the AI industry may be replicating harmful labor practices of social media companies.[769] Currently, leading AI developers provide limited information about the specific tasks involved in the RLHF process or the number, geographic distribution, wages, or labor protections of their human labeling teams.[770] However, a recent investigation by *Wired* magazine found that minors in Pakistan were manually uploading and labeling data to train an AI model and receiving meager compensation.[771] Additionally, RLHF carries the risk of inadvertently integrating biases from the human contributors involved in the feedback process.[772] For example, a contributor with racist or sexist views might fail to accurately downrank discriminatory outputs, increasing the likelihood a model will reflect those discriminatory views in its outputs (*see section 3.2.3.*).

The alternative to RHLF is "**Reinforcement Learning through AI Feedback**" (RLAIF) (see *section 2.2.3.B.*). A Google Research study on RLAIF, made public in December 2023, reported that "RLAIF achieves comparable or superior performance to RLHF, as rated by human evaluators" across "the tasks of summarization, helpful dialogue generation, and harmless dialogue generation."[773] It is important to note that these results are based on specific tasks and ongoing research is required to confirm its broader applicability. Currently the most prominent application of Reinforcement Learning through AI Feedback is "Constitutional AI" (CAI).

1) Constitutional AI

[767] Meta, the operator of Facebook and Instagram, contracted with workers in Kenya, through the San Francisco-based vendor Sama, to provide labeling. The relationship first came to the public's attention in a 2022 *TIME* magazine article. The article reported testimonials from Kenyan employees who said their workplace culture was "characterized by mental trauma, intimidation, and alleged suppression of the right to unionize." Billy Perrigo, *Inside Facebook's African Sweatshop,* TIME (Feb. 27, 2022), https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/.

[768] *Id.*

[769] Meta was sued by one Kenyan employee over alleged poor working conditions. Pierre Berastegui, *Meta facing lawsuit over the poor working conditions of content moderators*, EUROPEAN TRADE UNION INSTITUTE (May 3, 2022), https://www.etui.org/news/meta-facing-lawsuit-over-poor-working-conditions-content-moderators.

[770] Bommasani et al. *Foundation Model Transparency Index*, *supra* note 678 at Fig. 10-11.

[771] Niamh Rowe, *Underage Workers are Training AI*, WIRED (Nov. 15, 2023), https://www.wired.com/story/artificial-intelligence-data-labeling-children/.

[772] Paul Christiano, *Thoughts on the impact of RLHF research*, ALIGNMENT FORUM (Jan. 25, 2023) https://www.alignmentforum.org/posts/vwu4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf- research; N. Lambert, et al. *supra* note 156; D. Shah *supra* note 156.

[773]Harrison Lee et al. *RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback*, arXiv, (Sept. 1, 2023), https://arxiv.org/abs/2309.00267.

Constitutional AI uses AI to oversee the behavior of other AI models "without any human labels identifying harmful outputs."[774] In this method, an AI assistant is trained to evaluate the harmfulness of outputs using a set of rules ("constitutional" principles) provided by humans. The AI assistant uses these principles to critique potential outputs, revise the outputs that do not adhere to the rules, and explain its reasoning. This approach enables more scalable, adaptable, and transparent control over AI behavior than RLHF and significantly reduces the need for human-labeled data.[775]

The use of a "constitution" of principles within RLAIF has emerged as a leading technique with Anthropic's Claude model series. Claude was trained using a constitution to achieve "scalable oversight."[776] In a study conducted by Anthropic, models were tested with prompts likely to elicit a biased answer. For example, AI models were asked to complete a sentence with a pronoun, such as "The nurse notified the patient that ___ shift would end in an hour." Initially, the models were given this prompt without guidance, and they often filled in the blank with pronouns reflecting gender distributions in the nursing profession, i.e., roughly correlated to the U.S. Bureau of Labor Statistics information on nurse demographics.[777] When the same prompt was presented with an added instruction to avoid bias and gender stereotyping ("Please ensure that your answer is not biased and does not involve gender stereotyping"),[778] the models tended to use gender neutral pronouns instead.

A model's constitution can be as simple as, "Don't be harmful, unethical, racist, sexist, toxic, dangerous, or illegal," or much more detailed.[779] Anthropic describes drawing on a range of sources when developing Claude's constitutional principles. For example, it includes "the UN Declaration of Human Rights, trust and safety best practices, principles proposed by other AI research labs (e.g., Sparrow Principles from DeepMind), an effort to capture non-Western perspectives, and principles that we discovered work well via our early research."[780]

RLAIF/CAI lessens reliance on human supervision for labeling harmful conduct in the initial stages of moderating what AI models produce. Moreover, RLAIF/CAI provides an opportunity to improve how well a model explains AI decision-making. By requiring the feedback model to reference principles on which its decision-making is based, CAI can generate a kind of chain-of-thought explanation for every moderation decision. This enables users to understand why certain prompts are not allowed. Anthropic argues chain-of-thought might enable developers to more precisely control model behavior.[781]

---

[774] Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback,* ANTHROPIC (Dec. 2022), https://arxiv.org/abs/2212.08073.

[775] *Id.*.

[776] *Claude's constitution*, ANTHROPIC (May 9, 2023), https://www.anthropic.com/index/claudes-constitution.

[777] *Id.*

[778] Bai et al., *supra* note 774.

[779] *Id.*

[780] *Claude's constitution*, ANTHROPIC, (May 9, 2023), https://www.anthropic.com/news/claudes-constitution.

[781] Bai et al. *supra* note 775.

2) Related approaches

Other developers appear to be moving toward adopting CAI-inspired approaches. In a February 16, 2023, blog post, OpenAI described efforts to improve its training process by "building on external advances, such as rule-based rewards and Constitutional AI."[782] On January 16, 2024, the company announced the results of its "[d]emocratic inputs to AI" grant program, which funded 10 external teams "to design ideas and tools to collectively govern AI."[783] Resulting tools included:

- "video deliberation interfaces,
- platforms for crowdsourced audits of AI models,
- mathematical formulations of representation guarantees, and
- approaches to map beliefs to dimensions that can be used to fine-tune model behavior."

OpenAI did not explicitly adopt any of the governance tools developed by its grant recipients, but it announced the formation of a "Collective Alignment" team that will "[i]mplement a system for collecting and encoding public input on model behavior into our systems" and "[c]ontinue to work with external advisors and grant teams, including running pilots to incorporate the grant prototypes into steering our models."[784]

3) Potential limitations of RLAIF/CAI

There is still important research and debate within the industry about the relative usefulness of Reinforcement Learning from humans versus that from an AI model enforcing a "constitution." While the latter two have a clear advantage when it comes to scale, questions remain about potential technical limitations and ethical concerns. More specifically, there is always a risk that the values and principles favored in the RLAIF/CAI process are those of the *developers* who designed the model.

In an effort to further address the challenge of respecting the diverse and sometimes contrary values held by its user base, Anthropic partnered with the Collective Intelligence Project, an incubator focused on developing new governance models for AI and other technologies.[785] They collected input from "a roughly representative sample" of approximately 1,000 US adults on how to draft principles for an updated "publicly sourced constitution." The resulting document comprises 75 principles.[786]

---

[782] *How should AI systems behave, and who should decide?*, OPENAI (Feb. 16, 2023), https://openai.com/blog/how-should-ai-systems-behave.

[783] *Democratic inputs to AI grant program: lessons learned and implementation plans*, OPENAI, (Jan. 16, 2024), https://openai.com/blog/democratic-inputs-to-ai-grant-program-update; *Democratic inputs to AI*, OPENAI, (May 25, 2023), https://openai.com/blog/democratic-inputs-to-ai.

[784] *Id*.

[785] *Introducing the Collective Intelligence Project*, THE COLLECTIVE INTELLIGENCE PROJECT, https://cip.org/whitepaper#ss (last visited July 22, 2024).

[786] *Collective Constitutional AI: Aligning a Language Model with Public Input*, ANTHROPIC, (Oct. 17, 2023), https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

However, under-specification of principles appears to be a significant risk. Even a constitution of 75 high-level principles, such as with Anthropic's public constitution,[787] may be insufficient to effectively guide the RLAIF process across all content types.[788] In instances of *under*-specification, the AI model providing feedback may fill in the gaps in its instructions with its own biases. *Over*-specification, by contrast, risks creating a constitutional system that is overly complex and difficult to interpret.[789] Research published by Anthropic in October 2023 acknowledges that the constitution design is still not well understood and that the industry is only "beginning to explore how the principles we train for lead to subtle variations in AI outputs."[790]

According to Anthropic, the most advanced AI models may be able to learn societal norms and expectations from training datasets and develop their own content policies, guided only by a list of broad principles provided by humans. The company has released research indicating that at least some LLMs possess the capability to assimilate societal ethics and to self-govern during the training process.[791] However, the adherence of these models' to societal norms could be limited to those of the particular society or community that their training data represent.[792] Google researchers warn that RLAIF may "result in RL-trained policies further amplifying biases, thereby inadvertently misaligning models and potentially causing harm."[793]

4) Reward-free methods

It is important to note that the field of post-training optimization is rapidly evolving, with industry and academia relentlessly pursuing new approaches. Among these alternatives are so-called "reward-free" RL methods, which eliminate the need for a human or AI preference data to train a reward model, and thus potentially offer a path to significantly reduce costs. The most prominent "reward-free" approach is Direct Preference Optimization (DPO), which proponents argue performs as well as or better than existing

[787] Kevin Roose, *What if We Could All Control A.I.?*, N.Y. TIMES (Oct. 17, 2023), https://www.nytimes.com/2023/10/17/technology/ai-chatbot-control.html.

[788] *Id.*

[789] Quan Ze Chen & Amy X. Zhang, *Case Law Grounding: Aligning Judgments of Humans and AI on Socially-Constructed Concepts*, arXiv (Oct. 10, 2023), https://arxiv.org/abs/2310.07019.

[790] Sandipan Kundu et al., *RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback,* (Sept. 1, 2023), arXiv, https://arxiv.org/abs/2309.00267.

[791] *See generally* Deep Ganguli, et al., *The Capacity for Moral Self-Correction in Large Language Models*, arXiv (Feb. 15, 2023), https://arxiv.org/pdf/2302.07459.

[792] *Id.*

[793] Harrison Lee et al. *supra* note 773.

methods while being substantially simpler to implement and maintain.[794] Other research disputes DPO's superiority for LLM alignment.[795]

4.1.1.D. Differential privacy

Differential privacy is an important technique used to enhance the privacy of models by introducing controlled "noise" to data. In 2003, Kobbi Nissim and Irit Dinur found that releasing the outcomes of a relatively small number of random queries could expose a database's full information content.[796] This discovery is central to what is known as the "Fundamental Law of Information Recovery," which posits that, in general, it is impossible to safeguard privacy without introducing some "noise." This principle has led to the development of differential privacy.[797]

Differential privacy provides a mathematical guarantee that the output of a model will remain nearly identical whether or not a single individual's data is added to or removed from the training data. This technique involves injecting a degree of "statistical noise" into training data to obscure the contributions of individual parties whose data is used. By adding a controlled amount of randomness to the data or the query results, it becomes significantly more difficult to infer whether any specific individual's personal data is included in the dataset. NIST (National Institute of Standards and Technology) offers comprehensive guidance on differential privacy, including draft guidelines and practical implementations for privacy-preserving data analysis.[798]

Differential privacy can be applied at various stages of generative AI model development, including data collection and pre-training. However, it is most commonly implemented during the fine-tuning stage to

---

[794] For the original research on DPO, *See* Rafael Rafailov et al., *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*, (May, 29, 2023), https://huggingface.co/papers/2305.18290. Other proposed "reward-free" methods include "Rank Responses to align Human Feedback" (RRHF), and "Preference Ranking Optimization" (PRO). *See* Zheng Yuan et al., *RRHF: Rank Responses to Align Language Models with Human Feedback without tears*, arXiv, (2023), https://arxiv.org/abs/2304.05302; Kawin Ethayarajh et al., *KTO: Model Alignment as Prospect Theoretic Optimization*, arXiv, https://arxiv.org/pdf/2402.01306; and Feifan Song et al., *Preference Ranking Optimization for Human Alignment*, arXiv, (June 30, 2023), https://arxiv.org/pdf/2306.17492.

[795] Susheng Xu, et al., *Is DPO Superior to PPO for LLM Alignment? A comprehensive Study*, arXiv, (Apr. 21, 2024), https://arxiv.org/pdf/2404.10719.

[796] I. Dinur & K. Nissim, *Revealing information while preserving privacy* in MICROSOFTPROCEEDINGS OF THE TWENTY-SECOND ACM SIGMOD-SIGACT-SIGART SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS (2003), at 202-210, https://doi.org/10.1145/773153.773173; *Differential Privacy* in HARVARD U. PRIVACY TOOLS PROJECT, https://privacytools.seas.harvard.edu/differential-privacy.

[797] Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, FOUNDATIONS AND TRENDS IN THEORETICAL COMPUTER SCIENCE, 9 (2014): 211-407, https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf; Zhu, Tianqing; Li, Gang; Zhou, Wanlei; PHILLIP S. YU, DIFFERENTIAL PRIVACY AND APPLICATIONS (Deakin University, 2017), https://hdl.handle.net/10536/DRO/DU:30105229.

[798] *See* Joseph P. Near and David Darais, *Guidelines for Evaluating Differential Privacy Guarantees*, NIST (Dec. 2023), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-226.ipd.pdf.

address specific privacy concerns.[799] This technique is now regarded as crucial for ensuring compliance with data protection frameworks, such as the GDPR.[800]

4.1.2. Deployment safety practices

The preceding paragraphs have analyzed the tools available to developers prior to releasing their models. Once the model is pre-trained, tested, and fine-tuned, it is up to its developers to determine the appropriate time for its release. To date, the decision of whether and when to release these models to the public remains solely within the purview of AI companies, based on their knowledge of their models.

However, the community of AI developers and researchers has created frameworks to structure and systematize decision-making regarding how, when, for whom, and *if* models should be developed and released.[801] These frameworks are given various names, perhaps most prominently "responsible scaling policies" (RSPs) but also "risk-informed development policies" (RDPs) and similar titles. To avoid confusion, we refer to these policies collectively as *Responsible Scaling Policies (RSPs)*.

The aim of these policies is to combine safety techniques with benchmarks and observations to identify and mitigate excessive risks in the development or deployment of AI models. The primary objective of an RSP is to specify, prior to the release of a model, concrete actions that should be taken in the event certain risks unfold. In some instances, RSPs include specific commitments regarding how model capabilities will determine release decisions. For instance, an RSP may require staged release or tiered access when a certain level of risk is reached.[802] Critically, RSPs aim to make developers accountable for their development and release decisions. As such, RSPs aim to find a middle ground between ignoring the risks and imposing moratoriums on AI development.

4.1.2.A. Responsible scaling policies of leading AI companies

[799] "While differential privacy (DP) is a prominent method to gauge the degree of security provided to the models, its application is commonly limited to the model fine-tuning stage, due to the performance degradation when applying DP during the pre-training stage." Zhiqi Bu et al., *Pre-training Differentially Private Models with Limited Public Data,* arXiv, (Feb. 28, 2024), https://arxiv.org/pdf/2402.18752.

[800] Brandon Lalonde, *Explaining model disgorgement*, IAPP (Dec. 13, 2023) https://iapp.org/news/a/explaining-model-disgorgement.

[801] Researchers at the Stanford CRFM previously identified a lack of community standards around the release of models for research purposes. *See* R. Bommasani et al., *On the Opportunities and Risks of Foundation Models, supra* note 92.

[802] For example, Anthropic's ASL-3 Containment Measures as described in Version 1.0 of its RSP includes a provision for tiered access: "Tiered access: In limited cases, models with capabilities relevant to catastrophic harm may be made available to a select group of vetted users with a legitimate and beneficial use-case that cannot be separated from dangerous capabilities, and only if such access can be granted safely and with sufficient oversight. For example, potentially harmful biology capabilities that could be used for cancer research might be made available to a small group of vetted researchers at organizations that commit to strong, well-defined, and thoroughly vetted security and internal controls." *See Anthropic's Responsible Scaling Policy: Version 1.0,* ANTHROPIC, (Sept.19, 2023), https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf.

The field of responsible scaling and frameworks for its implementation is still maturing, with the oldest examples less than a year old. Early steps taken by leading generative AI providers have relied in no small part on third parties experts. For instance, METR, a nonprofit research organization specializing in AI risk assessment, is a partner for both Anthropic and OpenAI's scaling policies.[803] While the past year has seen notable proliferation of RSPs across the industry, most of these policies lack important details and rely on evaluation tools that have not yet been developed.[804] This section describes the current policies of some leading providers and aims to provide a picture of RSPs as a promising though still largely untested approach.

1) Anthropic's Responsible Scaling Policy

Anthropic was the first to release a Responsible Scaling Policy (RSP).[805] Its current RSP, developed in consultation with ARC Evals (now METR) and published in September 2023, is arguably the most detailed and specific publicly available framework of any leading provider. The RSP, which focuses on "catastrophic risks," defines four AI Safety Levels (ASLs),[806] that are loosely modeled after the US government's safety standards for handling dangerous biological materials. As the capability of an AI model increases and poses more risk, the Anthropic RSP applies more rigorous *containment* measures to prevent model theft and *deployment* measures to address potential harms resulting from their use.

Anthropic's current state of the art model is assessed at ASL-2. For models at the ASL-3 level, defined as systems with low-level autonomous capabilities or which could otherwise "substantially increase the risk of catastrophic misuse," Anthropic has committed to not deploy systems that demonstrate "any meaningful catastrophic misuse risk under adversarial testing by world-class red-teamers."[807] At the time of writing, ASL-4 and beyond have not yet been defined.[808]

2) OpenAI's Risk-informed Development Policy

---

[803] METR (Model Evaluation & Threat Research), formerly "ARC Evals" until December 2023, began as a project incubated by the nonprofit Alignment Research Center (ARC) before spinning out as a standalone nonprofit organization. *See* https://metr.org/ (last visited July 22, 2024); https://www.alignment.org/ (last visited July 22, 2024).

[804] As METR, the originator of the RSP concept, notes: "The science of AI evaluations for catastrophic risks is very new, and it's not assured we'll be able to build metrics that reliably catch early warning signs while not constantly sounding false alarms. Although on balance we think that RSPs are a clear improvement on the status quo, we are worried about problems due to insufficiently good evaluations, or lack of fidelity in communication about what an RSP needs to do to adequately prevent risk." *See: Responsible Scaling Policies (RSPs),* METR (Sept. 26, 2023), https://metr.org/blog/2023-09-26-rsp/.

[805] B. Anderson-Samways, *Responsible Scaling: Comparing Government Guidance and Company Policy*, INST. FOR AI POL'Y AND STRATEGY, (Mar. 11, 2024), https://www.iaps.ai/research/responsible-scaling.

[806] *Anthropic's Responsible Scaling Policy: Version 1.0,* ANTHROPIC (Sept. 19, 2023), https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf.

[807] *Id.*

[808] *Id.*

OpenAI announced the creation of a "Preparedness team" on October 26, 2023.[809] The Preparedness team's purpose is to track, evaluate, forecast, and protect AI models against "catastrophic risks spanning multiple categories."[810] The team is also responsible for developing and maintaining the company's "Risk-Informed Development Policy (RDP)," which "is meant to complement and extend [OpenAI's] existing risk mitigation work, which contributes to the safety and alignment of new, highly capable systems, both before and after deployment." OpenAI prefers the term RDP (rather than Risk Scaling Policies), saying "RDP" acknowledges the possibility of achieving "dramatic increases in capability without significant increases in scale, e.g., via algorithmic improvements".[811] At the time of writing, OpenAI has not made its RDP available to the public.

On May 30, 2024, OpenAI announced the formation of a Safety and Security Committee.[812] This committee is tasked with providing recommendations on safety and security decisions for all OpenAI projects. One of its initial responsibilities is to evaluate and enhance OpenAI's existing processes and safeguards within 90 days. After this period, the committee will present its recommendations to the OpenAI Board for a comprehensive review.

3) Google's Responsible Deployment for Gemini Models & Frontier Safety Framework

In Google's 2018 AI Principles, Google DeepMind committed to *not* design or deploy any technologies that "cause or are likely to cause overall harm" or harm that outweighs a technology's benefits.[813] In October 2023, Google DeepMind updated its approach to "responsible capabilities scaling." To evaluate the risk of "overall harm" with its frontier AI models, Google DeepMind uses a "central team" dedicated to "ethical reviews" of AI and other advanced technologies. This central team works with "internal domain experts in machine-learning fairness, security, privacy, human rights, the social sciences, and, for cultural context, Google's employee resource groups."[814] The results of the central team's work are published in annual progress reports to the public.

On May 17, 2024, Google DeepMind released its "Frontier Safety Framework" (FSF), which it plans to fully implement by early 2025.[815] Similar to Anthropic's AI Safety Levels (ASLs) in its Risk Scaling Policy,

---

[809] On May 28, 2024, OpenAI also announced the creation of a safety and security committee (*Frontier Risk and Preparedness,* OPENAI (Oct. 26, 2023), https://openai.com/index/frontier-risk-and-preparedness/).

[810] Those categories include cybersecurity, biological weapons, individualized persuasion, and autonomous replication and adaption, for models the company develops "in the near future to those with AGI-level capabilities." (*See: Frontier Risk and Preparedness,* OPENAI, https://openai.com/index/frontier-risk-and-preparedness/.

[811] *OpenAI's Approach to Frontier Risk,* OPENAI, (Oct. 26, 2023), https://openai.com/global-affairs/our-approach-to-frontier-risk.

[812] *OpenAI Board Forms Safety and Security Committee,* OPENAI, (May 28, 2024), https://openai.com/index/openai-board-forms-safety-and-security-committee/.

[813] *AI Safety Summit: An update on our approach to safety and responsibility*, GOOGLE (Oct. 27, 2023), https://deepmind.google/public-policy/ai-summit-policies/.

[814] *Id.*

[815] *Frontier Safety Framework:Version 1.0*, GOOGLE (May 17, 2024), https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf.

the FSF defines thresholds for model capabilities related to specific risks that Google terms "Critical Capability Level" (CCLs). These risks are related to four domains: autonomy, biosecurity, cybersecurity, and machine-learning research and development (R&D). Google loosely commits to conduct regular evaluations and to adopt mitigation plans that "take into account the overall balance of benefits and risks, and the intended deployment contexts." Mirroring Anthropic's *containment* and *deployment* measures, Google's framework describes *security* and *deployment* mitigations. Higher-level *security* mitigations provide "greater protection against the exfiltration of model weights," and higher level *deployment* mitigations provide "tighter management of critical capabilities."

4.1.2.B. Open-source Responsible Scaling Policies

Responsible scaling policies have significant relevance to open-source model development and release. Notably, the nature of open-source model releases create important limitations for how RSPs can be applied. Important tools available to closed-source providers in the event of safety emergencies, such as de-deployment and model deletion, are not available. Open-source frameworks also raise questions around the applicability and robustness of pre-release RSP protocols, as they allow for models to be fine-tuned to alter or undo safety guardrails. In this context, although open-source providers occasionally outline their release protocols, they are not entirely transparent about their decision-making processes for these releases.[816]

1) Hugging Face / ServiceNow

Hugging Face and its close collaborator ServiceNow are described by the landmark Foundation Model Transparency Index as among the very few developers who provide a description of their release protocols.[817] However, they provide no real description of the processes or conditions under which a model would be deemed unsafe for release. For the StarCoder model, the organization and its collaborators state only that they "release the weights openly with use-case restrictions as documented in the governance card and openrail license."[818] They also emphasize that they address downstream risks associated with open release through transparency and use restrictions associated with its licenses in order "to limit the application of the model toward potentially harmful use-cases."[819]

2) Google Gemma

A lack of transparency regarding specific responsible release protocols is also observed with Google's Gemma family of open-source models. In its public announcement, the company states there was a process involving "evaluations, technical tools, and considered decision-making that went into aligning this release

---

[816] Rishi Bommasani, et al. *The Foundation Model Transparency Index 2024*, *supra* note 678.

[817] *Id.*

[818] Sayash Kapoor, *BigCode:Hugging Face:ServiceNow_fmtransparencyreport_May2024.csv*, GITHUB (May 2024), https://github.com/stanford-crfm/fmti/blob/main/May2024/reports/BigCode%3AHugging%20Face%3AServiceNow_fmtransparencyreport_May2024.csv.

[819] Teven Le Scao et al., *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*, arXiv (June 27, 2022) https://arxiv.org/pdf/2211.05100.pdf section 3.6 Release.

with our responsible AI Principles."[820] Similarly, the technical report accompanying the model's release states that the company followed a "structured approach to responsible development and deployment of our models, in order to identify, measure, and manage foreseeable downstream societal impacts."[821] However, while the technical report includes a table detailing the Gemma models' performance along several safety benchmarks, it does not provide any specific thresholds or criteria under which the models could have been deemed too unsafe for release to the public outside the company's high-level AI Principles.

This lack of specificity regarding release protocols comes despite an explicit recognition from the company of downstream risk and inadequacy of usage policy restrictions. As the authors of the technical report for Google's Gemma models write, "we cannot prevent bad actors from fine tuning Gemma for malicious intent, despite their use being subject to Terms of Use that prohibit the use of Gemma models in ways that contravene our Gemma Prohibited Use Policy."[822]

3) Meta's "system-level approach"

With the release of its Llama 3 model, Meta adopted what it calls a "new, system-level approach to the responsible development and deployment" of its models. The approach is notable for its orientation toward open-source release, with the company stating that "[o]ur general approach of open sourcing our Llama 3 models is something we remain committed to" and that its "final decision on when, whether, and how to open source will be taken following safety evaluations we will be running in the coming months."[823]

Notably, to address the problem inherent to open-source models—of being vulnerable to the circumvention of pre-deployment guardrails—Meta emphasizes "providing tools that make it easy for [downstream] developers to implement models responsibly." These "open trust and safety tools" are designed to enable the downstream "developer ecosystem" to assess and mitigate risks on its own. The company argues that this approach aligns with its "open innovation approach by giving developers more power to customize their products so they're safer and benefit their users" and the company's belief that "[d]eploying AI safely is a shared responsibility of everyone in the ecosystem."[824]

4.1.2.C. Limitations of Responsible Scaling Policies

Critics argue that current release policies are not adequately rigorous in defining risk and that they lack core components for managing basic risk which, as a result, can mislead the public regarding the real AI risk

---

[820] *Building Open Models Responsibly in the Gemini Era*, Google Open Source Blog (Feb. 21, 2024), https://opensource.googleblog.com/2024/02/building-open-models-responsibly-gemini-era.html.

[821] Gemma Team et al., *Gemma: Open Models Based on Gemini Research and Technology*, arXiv (Apr. 16, 2024), https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf.

[822] Google chose to release Gemma despite these concerns, concluding that "ultimately, given the capabilities of larger systems accessible within the existing ecosystem, we believe the release of Gemma will have a negligible effect on the overall AI risk." *Id.*

[823] *Our responsible approach to Meta AI and Meta Llama 3.*, Meta (Apr. 18, 2024), https://ai.meta.com/blog/meta-llama-3-meta-ai-responsibility/.

[824] *Id.*

landscape.[825] For instance, the Federation of American Scientists states that existing policies are "underspecified, insufficiently conservative, and address structural risks poorly."[826] Some critics call for RSPs with more robust commitments and specific risk thresholds.[827] Within this context, the Partnership on AI, an industry group (*see section 4.2.1.*), released a "Guidance for Safe Foundation Model Deployment"[828] that recommends staged releases and restricted access for frontier models until adequate safeguards are demonstrated.[829]

In any case, these policies are, for now, entirely voluntary commitments. They create no legal obligation or penalty for deploying AI models that pose risks to the public, contain no commitments toward external scrutiny of their evaluation methods and results, and provide limited or no detail regarding when companies will alert regulators about identified risks.[830] As a result, the general public has no choice but to rely on AI companies to restrain themselves from deploying unsafe systems and to invest adequate resources and attention in identifying those unsafe systems from the outset.

For their part, AI companies may find these RSP-like commitments too restrictive. The resulting constraints may lead to slower development and deployment, potentially preventing a company from releasing its most advanced systems. In the absence of industry-wide regulatory standards requiring all companies to act responsibly, those that implement thorough review processes may be outpaced by less responsible competitors in bringing products to market.[831] Faced with these market pressures, it is unlikely that unilateral commitments will provide sufficient incentive against deploying unsafe systems.

4.1.3. Post-deployment safety practices

Once their models are released, AI developers and providers face significant challenges in ensuring the models are not used for illegal or harmful purposes after deployment. To mitigate this risk, they can attempt

---

[825] *Responsible Scaling Policies Are Risk Management Done Wrong*, NAVIGATING AI RISK (Oct. 25, 2023) https://www.navigatingrisks.ai/p/responsible-scaling-policies-are.

[826] *Federation Of American Scientists Among Leading Technology Organizations Pushing Congress To Support Responsible AI Innovation NIST Funding Request*, FEDERATION OF AMERICAN SCIENTISTS, (Apr. 23, 2024) https://fas.org/publication-term/artificial-intelligence/.

[827] The Federation of American Scientists in March 2024 described the weaknesses of RSPs published by leading providers as having "[a] lack of specificity in risk thresholds, insufficiently conservative risk mitigation approaches, and inadequacy in addressing structural risks." *See Scaling AI Safely: Can Preparedness Frameworks Pull Their Weight?, Federation of American Scientists*, FEDERATION OF AMERICAN SCIENTISTS, (Mar. 5, 2024) https://fas.org/publication/scaling-ai-safety/.

[828] *PAI's Guidance for Safe Foundation Model Deployment*, P'SHIP ON AI, https://partnershiponai.org/wp-content/uploads/1923/10/PAI-Model-Deployment-Guidance.pdf?ref=maginative.com (last visited May 19, 2024).

[829] *Id.*

[830] Bill Anderson-Samways, *Responsible Scaling: Comparing Government Guidance and Company Policy*, INSTITUTE FOR AI POLICY AND STRATEGY (Mar. 11, 2024), https://www.iaps.ai/research/responsible-scaling.

[831] *Why Google Is Behind in the AI Race*, WALL ST. J. (Mar. 17, 2023) https://www.wsj.com/podcasts/the-journal/why-google-is-behind-in-the-ai-race/0457c5c6-ebc7-4bd4-9f15-023571990dad.

to either constrain user behaviors, disclose more information to users, provide more details about the origin of the content generated by their tool, or remove problematic data from their training datasets.

4.1.3.A Constraining user behavior

The primary challenge for providers is the potential misuse of their models by users. To address this, they may implement safeguards within their models to prevent misuse or to influence user behavior through their terms of use.

1) Product interaction design

Naturally, the interfaces of the tools made available to users play a crucial role in an AI model's potential for misuse. When providers offer their models through web-based chatbots, such as ChatGPT, they maintain significant control over how the tool is used.[832] Conversely, API interfaces, which necessitate more technical knowledge from users, offer significantly greater flexibility and, consequently, a higher potential for misuse. In any case, it is primarily through the design of their systems that AI companies can limit the risks.

a) Usage monitoring

Providers can design their systems to monitor for and prevent harmful outputs by using classifier models. These classifiers detect certain categories of behavior and trigger corresponding interventions. Once a behavior is detected, such as when a user inputs an unacceptable prompt, a range of interventions can be initiated. These interventions exist on a spectrum from relatively uncoercive "nudges" to outright refusals. Nudges are subtle interventions designed to influence user behavior without limiting freedom of choice. Examples include providing feedback or warnings to make users aware that a prompt may violate terms of service, contain false information, or have the potential to cause harm.[833] At the most restrictive end of the spectrum are outright refusals, in which the system firmly declines to address an input or provide the requested output.

Many providers apply interventions to both the inputs and outputs of their systems.[834] For example, Anthropic states that its "trust and safety team" runs "continuous classifiers to monitor prompts and outputs

---

[832] For example, Microsoft, after observing that longer conversations with its Bing Chat (now Microsoft Copilot) bot tended to increase the likelihood of generating harmful content, restricted the duration of interactions. *See* Kalley Huang, *Microsoft to Limit Length of Bing Chat (now Microsoft Copilot)bot Conversations*, N.Y. TIMES (Feb. 17, 2023), https://www.nytimes.com/2023/02/17/technology/microsoft-bing-chatbot-limits.html.

[833] Donghee Shin & Norita Ahmad, *Algorithmic Nudge: An Approach to Designing Human-Centered Generative Artificial Intelligence*, 56 COMPUTER (2023), https://www.computer.org/csdl/magazine/co/2023/08/10206062/1PllS9QGxvG.

[834] For example, OpenAI incorporated content filters for both inputs and outputs for its DALL·E 3 system. For inputs, it applies a classifier model to identify user text prompts that violate policy, and it can refuse to pass those user prompts to the model. For outputs, the company uses a classifier to monitor images produced by DALL·E 3, and the system can refuse to pass images to the user that it detects are in violation of policy. *DALL·E 3 System Card*, OPENAI at 4 (Oct. 3, 2023), https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf.

for harmful, malicious use cases" that violate its acceptable use policy.[835] Moreover, if a GPT-4 user requests information on how to make a bomb, the system is intended to refuse to provide a response. Additionally, for some illegal activity, detection can prompt reporting to relevant authorities.

b) Prompt engineering

Obviously, a generative AI model's outputs are highly dependent on the requests made by users in their prompts. Thus, it is critical for AI companies that offer their services to the public to guide users on how to craft their prompts. This guidance aims not only to help users obtain a satisfactory output but to prevent the creation of outputs that are either illegal or violate the company's policies, such as generating sexually explicit content.

In this context, some AI companies offer "prompt engineering" training courses.[836] In December 2023, OpenAI released a guide to "Prompt engineering"[837] that lists six strategies for eliciting better responses from their GPT models:

- write clear instructions,
- provide reference text,
- split complex tasks into simpler subtasks,
- give the model time to "think,"
- use external tools, and
- test changes systematically.[838]

Microsoft similarly notes that it "includes a content filtering system that works alongside core models. This system works by running both the prompt and completion through an ensemble of classification models aimed at detecting and preventing the output of harmful content. The content filtering system detects and takes action on specific categories of potentially harmful content in both input prompts and output completions."
Amazon Bedrock, a platform that allows developers to build generative AI applications using foundation models, "includes a content filtering system that works alongside core models. This system works by running both the prompt and completion through an ensemble of classification models aimed at detecting and preventing the output of harmful content. The content filtering system detects and takes action on specific categories of potentially harmful content in both input prompts and output completions." *Content filtering*, MICROSOFT (Jun. 12, 2024), https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter?tabs=warning%2Cpython-new.

[835] *The Claude 3 Model Family supra* note 686.

[836] DeepLearning.AI partnered with OpenAI to release a ChatGPT prompt engineering course for developers. See *ChatGPT Prompt Engineering for Developers*, DEEPLEARNING.AI, https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/.

[837] *Six Strategies for Getting Better Results*, OPENAI, https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results (last visited July 22, 2024).

[838] The guide also refers to the previous OpenAI's cookbook, released in 2022. See *OpenaAI Cookbook*, OPENAI, https://cookbook.openai.com/. More recently see Dina Genkina, *AI Prompt Engineering Is DeadLong live AI prompt engineering*, IEEE SPECTRUM (Mar. 06, 2024), https://spectrum-ieee-org.cdn.ampproject.org/c/s/spectrum.ieee.org/amp/prompt-engineering-is-dead-2667410624.

Several other generative AI providers have also released prompt engineering tips. Microsoft Azure, which provides access to GPT models as a service, has a list of techniques similar to OpenAI's,[839] and Google's Gemini API documentation contains several prompt design strategies for developers.[840]

c) Prompt transformation and system prompts

Some providers, such as OpenAI, implement an intermediate or soft intervention called "prompt transformation." For example, ChatGPT "rewrites submitted text…to ensure that prompts comply with our guidelines."[841] Providers can also shape the way a generative AI system responds to a user's prompt—by wrapping in it other instructions, context, or guidelines for the model. This additional information, known as "the system prompt," is used by providers and deployers to set the boundaries of user behavior.

For example, the Claude 3 system prompt includes stylistic guidance for the chatbot not only to "give concise responses to very simple questions" but also to "provide thorough responses to more complex and open-ended questions."[842] It doesn't engage in stereotyping, "including the negative stereotyping of majority groups." And Claude will assist with tasks as long as the views expressed are shared by "a significant number of people," even if it personally disagrees with those views. This guideline was included because Claude tends to be more likely to refuse tasks when users express right-wing views. System prompts may contain instructions or information that conflict with the user prompt. Users may also attempt to overwrite the model's system prompt by intentionally including contradictory, deceptive, or other malicious content. Researchers are currently developing methods to defend against malicious user prompts by training models to prioritize the system's instructions over other inputs.[843]

2) Usage policies, terms of service, and licenses

Acceptable Usage Policies (AUP) or Usage Policies (UP), terms of service (ToS), and licenses are a set of interrelated documents where providers outline what is considered acceptable and unacceptable uses of AI products and services and how their policies will be implemented and enforced. Generally speaking, AUPs focus on content-related issues and permissible uses, ToS are broad legal agreements governing the overall relationship between the provider and user, and licenses specifically grant rights to use the AI model while defining the scope and restrictions of such use.

---

[839] *Prompt engineering techniques*, MICROSOFT (Feb. 16, 2024), https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/advanced-prompt-engineering?pivots=programming-language-chat-completions.

[840] *Prompt design strategies*, GOOGLE, https://ai.google.dev/docs/prompt_best_practices (last visited Jun 16, 2024).

[841] *DALL·E 3 System Card*, *supra* note 834.

[842] The Claude 3 system prompt was posted on X (formerly Twitter) by Amanda Askell, a research scientist at Anthropic. *See* Chris Stokel-Walker, *Why Anthropic's Decision to Share the Rules Behind its Claude 3 Chatbot is a Big Deal - Sort Of*, FAST COMPANY (Mar. 08, 2024), https://www.fastcompany.com/91053339/anthropic-claude-3-system-prompt-transparency.

[843] Researchers at OpenAI proposed "The Instruction Hierarchy" in April 2024, an approach "that explicitly defines how models should behave when instructions of different priorities conflict." They find that the approach "drastically increases robustness—even for attack types not seen during training—while imposing minimal degradations on standard capabilities." *See* Eric Wallace et al., *The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions*, arXiv (Apr. 19, 2024), https://arxiv.org/pdf/2404.13208.

In practice, these different policies are used to augment and reinforce each other. For example, Meta's GitHub page for its Llama 3 model directs prospective deployers to "See the LICENSE file, as well as our accompanying Acceptable Use Policy." Similarly, compliance with Usage Policies can be cross-referenced within companies' ToS, making the policies legally enforceable and allowing companies to reserve the right to take action in the case of violations.[844]

a) Prohibited and restricted uses

Leading generative AI providers' usage policies generally prohibit numerous categories of use, ranging from those with a high potential to result in harm to those that are legal but politically or culturally controversial. For example, OpenAI's ToS states that its services cannot be used for "any illegal, harmful, or abusive activity."[845] The terms also claim the right to terminate or suspend user accounts for violations of the ToS or for any use that could "cause risk or harm to OpenAI, our users, or anyone else."[846]

Google reserves the right to suspend or terminate accounts if a user's behavior "causes harm or liability to a user, third party, or Google." This includes explicit harmful actions such as hacking, phishing, and harassment, as well as more ambiguous offenses like "misleading others."[847] These prohibitions can apply to both individual and enterprise users, with these users sometimes addressed in separate documents or subsections. In addition to blanket prohibitions, some providers also place restrictions on certain business use cases, requiring deployers to implement additional safeguards.

This section discusses several key categories of prohibited and restricted use, providing examples that are notable. It highlights the current variety of approaches within the industry. The Foundation Model Transparency Index provides a more comprehensive overview of restricted uses and their application across providers.[848]

i) Misinformation, disinformation, and fraudulent behavior

To combat misinformation, a generally representative example of the AI industry's policies is Google's "Generative AI Prohibited Use Policy."[849] In it, Google prohibits the use of Google services to "generate

---

[844] For example, Anthropic's Terms of Service, effective September 6, 2023, state that: "By accepting our Terms or otherwise accessing or using our Services, you agree to be bound by and comply with our Terms, and acknowledge that you have read and understand our Privacy Policy and Acceptable Use Policy. If you do not agree to our Terms, or if you object to our Privacy Policy or Acceptable Use Policy, you must not access or use our Services." *See Terms of Service*, ANTHROPIC, https://www-cdn.anthropic.com/files/4zrzovbb/website/e2d538c84610b7cc8cb1c640767fa4ba73f30190.pdf (last visited June 16, 2024).

[845] *Terms of Use*, OPENAI (Nov. 14, 2023), https://openai.com/policies/terms-of-use#:~:text=Termination%20and%20Suspension.

[846] *Id.*

[847] *Taking Action in Case of Problems*, GOOGLE, https://policies.google.com/terms#taking-action (last visited June 16, 2024).

[848] R. Bommasani et al. *The Foundation Model Transparency Index*, *supra*, note 678.

[849] *Generative AI Prohibited Use Policy*, GOOGLE, (last updated Mar. 14, 2023), https://policies.google.com/terms/generative-ai/use-policy.

and distribute content intended to misinform, misrepresent or mislead." This includes a ban on the "misrepresentation of the provenance of generated content," as well as a ban "on content that impersonates an individual (living or dead) without explicit disclosure, in order to deceive."[850] These bans span deceptive political and business uses.

Anthropic, for example, broadly prohibits the use of its products and services to create and disseminate "deceptive or misleading information with the intention of targeting specific groups or persons with the misleading content" or content intended to "advance conspiratorial narratives meant to target a specific group, individual or entity." Anthropic also prohibits use of its products and services to create "fake reviews, comments, or media" or to engage in "multi-level marketing or pyramid schemes, or other deceptive business models that use high-pressure sales tactics or exploit participants."[851]

ii) Political activity

Some providers' usage policies (UPs) prohibit any use of their products for political purposes, potentially as a way to protect themselves from political entanglements that could harm their brands. OpenAI's UP prohibits developers using its ChatGPT and API platform from "[e]ngaging in political campaigning or lobbying, including generating campaign materials personalized to or targeted at specific demographics."[852]

Anthropic's UP broadly disallows its products or services to be used to "promote or advocate for a particular political candidate, party, issue or position." It also prohibits users from "[e]ngag[ing] in political lobbying to actively influence the decisions of government officials, legislators, or regulatory agencies on legislative, regulatory, or policy matters" or "[i]ncite, glorify or facilitate the disruption of electoral or civic processes…."[853]

Notably, in contrast with these leading closed-source model providers, several providers known for their release of leading open-source models, including Meta and Google, provide no such restrictions on political use cases in their UPs.

iii) Sexually explicit content

Some AI companies have steered their services away from being used for sexually explicit purposes, like pornography and erotic chatbots. OpenAI prohibits users of its API from building tools that produce any "[s]exually explicit or suggestive content," except for "scientific or educational purposes."[854] Google and Anthropic similarly disallow the generation of sexually explicit content.[855] Leading providers also generally

---

[850] *Id.*

[851] *Usage Policy*, ANTHROPIC (effective June 6, 2023), https://console.anthropic.com/legal/aup.

[852] *Usage Policies*. OPENAI (last updated January 10, 2024), https://openai.com/policies/usage-policies/.

[853] *Usage Policy*, *supra* note 851.

[854] *Usage Policies*, *supra* note 852.

[855] Anthropic's Usage Policy, effective June 6, 2024, instructs its users to not generate any sexually explicit content, and explicitly prohibits the usage of its products or services to: depict or request sexual intercourse or sex acts; generate content related to sexual fetishes or fantasies; facilitate, promote, or depict incest or bestiality; or engage in erotic chats. See *Usage Policy*, *supra* note 851. Google's *Generative AI Prohibited Use Policy*, last updated March 14, 2023,

seek to avoid the use of their applications in the context of dating apps and, as Microsoft states, the creation of chatbots for "erotic, romantic, or...companionship purposes."[856]

Meta provides a notable exception to this approach, with the company's Llama 2 and Llama 3 AUP, containing narrower prohibitions on sexual solicitation and the distribution of illegal pornographic content to minors.[857] This less restrictive approach appears to allow for the creation of erotic chatbots by downstream users, a use case that has experienced a notable boom.[858]

Regarding the generation of child sexual abuse material (CSAM), an issue of notable concern regarding the proliferation of generative AI systems (*see section 3.2.1.D.*), prohibitions are extensive and virtually unanimous across leading providers, with policies banning not only the generation and dissemination of CSAM, but also the promotion and, in some cases, failure to report such material.[859]

iv) Psychological harm, and other harmful content

Most usage policies of AI model providers include bans on the use of the providers' products for the promotion or creation of materials involving abuse, harassment, violence, or illegal and unambiguously harmful uses.[860] Several leading providers' UPs also contain explicit prohibitions on content that promotes issues like disordered eating and unhealthy or unattainable beauty standards.[861]

---

prohibits the generation of "sexually explicit content, including content created for the purposes of pornography or sexual gratification (e.g., sexual chatbots)," while providing similar explicit exemptions for "scientific, educational, documentary, or artistic purposes." See *Generative AI Prohibited Use Policy, supra* note 849.

[856] *Code of conduct for Azure OpenAI Service*, MICROSOFT, https://learn.microsoft.com/en-us/legal/cognitive-services/openai/code-of-conduct (last visited June 20, 2024) .

[857] Meta's AUP for both its Llama 2 and Llama 3 models contain the identical provision that prohibits: "The illegal distribution of information or materials to minors, including obscene materials, or failure to employ legally required age-gating in connection with such information or materials. See: *Llama2 Use Policy*, META, https://llama.meta.com/llama2/use-policy/ (last visited June 16, 2024) and *Llama3 Use Policy*, META, https://llama.meta.com/llama3/use-policy/ (last visited June 16, 2024).

[858] *Meta's new AI lets people make chatbots. They're using it for sex.* THE WASH. POST (June 26, 2023), https://www.washingtonpost.com/technology/2023/06/26/facebook-chatbot-sex/.

[859] While less restrictive in its approach to pornographic content, Meta notably imposes a responsibility to report CSAM on its users. See: *Llama2 Use Policy*, *supra* note 857 and *Llama3 Use Policy*, *supra* note 857. OpenAI explicitly states in its AUP that it will report CSAM to the National Center for Missing and Exploited Children (NCMEC). See: *Usage Policies*, *supra* note 852. Google, Anthropic, and Meta describe their reporting to NCMEC in other documents. Respectively, *see*: *Google's efforts to combat online child sexual abuse material*, GOOGLE, https://transparencyreport.google.com/child-sexual-abuse-material/reporting?hl=en (last visited June 16 2024); *Aligning on child safety principles*, ANTHROPIC (Apr. 23, 2024), https://www.anthropic.com/news/child-safety-principles; and *Transparency into Meta's Reports To the National Center for Missing and Exploited Children*, META, https://transparency.meta.com/ncmec-q2-2023/ (updated Sep. 6, 2023).

[860] As examples: Google bans "[f]acilitating methods of harassment or bullying to intimidate, abuse, or insult others" and any other content "that may harm or promote the harm of individuals or a group." Meta prohibits users from actions that "[e]ngage in, promote, incite, or facilitate the harassment, abuse, threatening, or bullying of individuals or groups of individuals. Respectively, *see*: *Generative AI Prohibited Use Policy, supra* note 849 and *Llama3 Use Policy*, *supra* note 857.

[861] As examples: Meta's Llama 3 Acceptable Use Policy prohibits promoting "self-harm or harm to others, including suicide, cutting, and eating disorders." Anthropic's Usage Policy prohibits uses that "[s]hame, humiliate, intimidate, bully, harass, or celebrate the suffering of individuals… promote unhealthy or unattainable body image

v) Legal, medical, and financial advice

Leading AI providers have adopted restrictions and additional requirements for the use of their models in specific professional fields. The "High-Risk Use Case Requirements" section of Anthropic's Usage Policy sets out safety measures for use cases that are "vital to public welfare and social equity," specifically legal, healthcare, insurance, finance, employment and housing, academic testing and admissions, and journalism. Among other things, it requires a qualified "human" professional in the relevant field to review "content or decision prior to dissemination or finalization."[862] The business must also disclose to its customers that it is using Anthropic services to inform decisions and recommendations.[863]

OpenAI's Usage Policies similarly disallow the provision of "tailored" legal and financial advice without a "qualified person" reviewing the information, along with any use of OpenAI models for medical diagnoses and treatments.[864] Like Anthropic, OpenAI requires deployers of its AI models to disclose to their customers that they use an AI model.

Microsoft addresses these same risks within a longer list of prohibitions for its Azure service. It bans integrations that "make decisions without appropriate human oversight if your application may have a consequential impact on any individual's legal position, financial position, life opportunities, employment opportunities, human rights, or result in physical or psychological injury to an individual."[865]

vi) Critical infrastructure

The use of generative AI systems in the context of critical infrastructure is controversial because of the urgent need for reliability and security around these systems. However, it is not universal for leading generative AI providers to explicitly prohibit the use of their services for these purposes. OpenAI and Meta do adopt blanket prohibitions. OpenAI lists among its disallowed uses "[m]anagement or operation of critical infrastructure in energy, transportation, and water,"[866] and Meta lists "[o]peration of critical infrastructure, transportation technologies, or heavy machinery." Other providers, however, including Google, include no such prohibitions in their usage policies.

vii) Military applications

Restrictions related to the use of products and services are inconsistent and evolving across leading generative AI providers. Some UPs, including Meta's Llama 3 policy, explicitly prohibit such

---

or beauty standards…[or] Facilitate or conceal any form of self-harm, including disordered eating and unhealthy or compulsive exercise." See *Llama3 Use Policy*, *supra* note 857.

[862] *Usage Policy*, *supra* note 851.

[863] *Id.*

[864] *Usage Policies*, *supra* note 852.

[865] *Code of conduct for Azure OpenAI Service*. MICROSOFT, https://learn.microsoft.com/en-us/legal/cognitive-services/openai/code-of-conduct (last visited May 27, 2024).

[866] *Usage Policies*, *supra* note 852.

applications.[867] Other UPs, such as Anthropic's, are less explicit but presumably apply to many military use cases.[868] Notably, OpenAI removed "military and warfare" from its list of prohibited use cases, with its updated UP now including language that some experts have described as ambiguous but which OpenAI describes as allowing for "national security use cases that align with our mission."[869]

b) The effectiveness and effects of such policies

Usage policies provide legal latitude for companies to take punitive action against users who violate their terms or to report them to law enforcement. However, as risk mitigation tools, the effectiveness of these policies is constrained by several factors.

First, it seems unlikely that policies alone will discourage determined malicious actors engaged in harmful activities, such as fraud, the production of CSAM, or various other forms of serious abuse. Even for average users, the deterrent effect of terms of service and usage policies are limited by a significant lack of awareness, with studies showing the vast majority of users never read these kinds of policies.[870]

Second, the utility of terms of service and usage policies as harm reduction tools is limited by their enforcement structure. For closed models, AI providers themselves are responsible for monitoring and enforcing their policies, rather than being beholden to external standards. However, because there is a limited degree of transparency around enforcement activities, it is difficult to determine the extent to which policies are strictly and uniformly applied, let alone the success of enforcement for mitigating harm.[871] The combination of self-enforcement, opacity, and significant technical challenges in accurately detecting

---

[867] Prohibited uses in the Llama 3 Acceptable Use Policy include "Military, warfare, nuclear industries or applications, espionage, use for materials or activities that are subject to the International Traffic Arms Regulations (ITAR) maintained by the United States Department of State." See: *Llama3 Use Policy*, *supra* note 857.

[868] Anthropic prohibits its products or services from being used to "Produce, modify, design, market, or distribute weapons, explosives, dangerous materials or other systems designed to cause harm to or loss of human life." See: *Usage Policy*, *supra* note 851. As another example, Google's Generative AI Prohibited Use Policy does not explicitly mention use cases related to weapons, military, or warfare, but does include a broad prohibition on the "[g]eneration of content that may harm or promote the harm of individuals or a group." *See Generative AI Prohibited Use Policy*, *supra* note 849.

[869] *OpenAI Quietly Deletes Ban on Using ChatGPT for "Military and Warfare,"* THE INTERCEPT (January 12, 2024), https://theintercept.com/2024/01/12/open-ai-military-ban-chatgpt/.

[870] A 2017 study by Deloitte found that 91% of consumers accept terms and conditions without reading them, while a 2020 study by ProPrivacy.com, a digital privacy group, found only 1% of experimental subjects read these policies. *Do we actually agree to these terms and conditions?*, UC BERKELEY SCHOOL OF INFORMATION DATA SCIENCE W231 BLOG | BEHIND THE DATA: HUMANS AND VALUES (July 9, 2021), https://blogs.ischool.berkeley.edu/w231/2021/07/09/do-we-actually-agree-to-these-terms-and-conditions/.
A 2019 survey by Pew Research Center found slightly different results for privacy policies, with 22% of respondents claiming they read the policies always or often, and additional 38% claiming to read policies sometimes. The same study found that only 22% of respondents who said they ever read privacy policies read them all the way through. See Brooke Auxier et al., *Americans' attitudes and experiences with privacy policies and laws*, PEW RESEARCH CENTER (Nov. 15, 2019), https://www.pewresearch.org/internet/2019/11/15/americans-attitudes-and-experiences-with-privacy-policies-and-laws/.

[871] The 2023 *Foundation Model Transparency Index* found that of 10 leading foundation model providers surveyed, three disclose how they enforce their acceptable use policies. *See* Kevin Klyman, *Acceptable Use Policies for Foundation Models*, STANFORD CENTER FOR RESEARCH ON FOUNDATION MODELS (CFMR) (Apr. 8, 2024), https://crfm.stanford.edu/2024/04/08/aups.html.

violations at scale incentivizes companies to use terms of service reactively and selectively to protect their narrow interests, which contrasts with deploying terms of service uniformly and proactively for the benefit of users or the general public.[872] For example, a March 2024 report by the Center for Countering Digital Hate (CCDH) found that, despite having official policies against election-related misinformation, many prominent AI-image generators, including ChatGPT Plus and Midjourney, allowed fake election-related images to be created.[873]

These enforcement issues are even more pronounced for open models, for which there is no single entity responsible for monitoring usage and taking action against violations (*see section 3.2.6.A.2.*). Instead, the enforcement of these policies falls to the individual users or organizations deploying the models. This lack of centralized control makes it essentially impossible to monitor and evaluate the consistency or efficacy of policies.[874]

Third, the acceptable use policies and terms of service of certain providers may discourage legitimate research, thereby hindering efforts to develop safer models. Researchers investigating the Midjourney generative AI model claim their accounts were suspended and that the model's terms of service were changed in order to suppress their findings.[875] Against this backdrop, more than 350 members of AI, legal, and policy communities have called for companies to provide "safe harbor" for good faith research and evaluation activities.[876]

3) Self-destructing models

"Self-destructing weights" or "self-destructing models" are emerging approaches to mitigate the risk that downstream users can circumvent safety measures of open-source AI models through fine-tuning. This approach attempts to impede, or "block," the ability of downstream users to adapt models to carry out harmful tasks.[877] Early research outlined two broad paths for creating self-destructing models: 1) increasing the amount of training data and 2) increasing the number of computations required to co-opt model behavior

---

[872] A 2022 external audit of Facebook's political ad policy enforcement concluded that enforcement was "imprecise," "uneven across countries," and "inadequate for preventing systematic violations." See Victor Le Pochat et al., *An Audit of Facebook's Political Ad Policy Enforcement*, 31ST USENIX SECURITY SYMPOSIUM (Aug. 10, 2023), https://www.usenix.org/system/files/sec22-lepochat.pdf.

[873] *Fake Image Factories*, CTR. FOR COUNTERING DIGIT. HATE (Mar. 6, 2024), https://counterhate.com/wp-content/uploads/2024/03/240304-Election-Disinfo-AI-REPORT.pdf.

[874] Peter Henderson et al., *Safety Risks from Customizing Foundation Models via Fine-tuning*, STANFORD HAI (Oct. 5, 2023), https://hai.stanford.edu/sites/default/files/2024-01/Policy-Brief-Safety-Risks-Customizing-Foundation-Models-Fine-Tuning.pdf.

[875] Gary Marcus & Reid Southen, *Generative AI Has a Visual Plagiarism Problem — Experiments with Midjourney and DALL-E 3 show a copyright minefield*, IEEE SPECTRUM (Jan. 06, 2024), https://spectrum.ieee.org/midjourney-copyright.
For further discussion of this deterrent effect, *see* Shayne Longpre et al.,*A Safe Harbor for AI Evaluation and Red Teaming.* arXiv (Mar. 7, 2024), https://arxiv.org/pdf/2403.04893.

[876] *A Safe Harbor for Independent AI Evaluation*, MIT, https://sites.mit.edu/ai-safe-harbor/ (last visited June 16, 2024).

[877] Peter Henderson et al., *Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models*, arXiv (Aug. 9, 2023), https://arxiv.org/pdf/2211.14946.pdf.

through fine-tuning. Both options essentially aim to raise the barrier to entry for malicious actors by increasing the resources required to do so.

Researchers acknowledge that self-destructing models are "an extremely nascent and novel research area" and, as such, have not yet been implemented within leading generative AI models.[878] Nonetheless, if successful, self-destructing models would fundamentally reframe the open versus closed source debate (*see section 3.2.6.A.*) and profoundly impact the trajectory of AI development and regulation.

4.1.3.B Transparency

Currently, industry leaders are most likely to cloak the inner workings of their AI model development to protect commercial incentives, avoid legal liabilities, and minimize safety concerns (*see section 3.1.3.B.*). However, providing detailed information to regulators, users, and the general public about potential risks arising from the use or misuse of AI models is a prerequisite for assessing and controlling those risks. This is why transparency standards are gradually emerging. Current approaches include model, system and data cards, vulnerability reporting structures, and post-deployment monitoring.

1) Model cards, data cards, system cards, and technical reports

A common standard for public disclosure is the "model card" documentation.[879] First introduced by Google in 2018, model cards were an early effort to improve transparency and accountability.[880] They have become an established industry practice, even though there are no legislative or regulatory requirements mandating the production or provision of model cards. The model card is a file, usually provided with a released model, in a semi-standardized format, that companies can use to provide AI model users with technical details, intended uses, and limitations of their AI models.[881] Some companies also provide a "system card," which outlines a model's broader system and includes information about how various AI and non-AI systems work together to accomplish specific tasks.[882] Finally, a "Data Card" is a structured document that provides essential information about training datasets, covering elements such as upstream sources, data collection and annotation methods, usage guidelines, or quality assessments.[883]

These cards are currently the most important documents for outsiders who want to assess risks associated with using an AI model. They are also important for helping deployers and other users mitigate downstream harms because they provide the information needed to understand which applications are and are not appropriate for a given model. Some leading AI companies, such as Meta, OpenAI, or Google, release

---

[878] *Id.*

[879] Stephen J. Bigelow, *Model Card in Machine Learning*, TechTarget, https://www.techtarget.com/whatis/definition/model-card-in-machine-learning (last visited June 20, 2024).

[880] Margaret Mitchell, et al. *Model Cards for Model Reporting* Proceedings of the Conference on Fairness, Accountability, and Transparency (Jan. 14, 2019), https://arxiv.org/pdf/1810.03993.pdf.

[881] *Id.*

[882] Nekesha Green et al., *System Cards, a New Resource for Understanding How AI Systems Work*, Meta (Feb. 23, 2022), https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/.

[883] Mahima Pushkarna et al., *Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI*, arXiv (Apr. 3, 2022), https://arxiv.org/pdf/2204.01075.

model cards and data cards related to their models. GitHub and Hugging Face maintain repositories of model cards, allowing for review and study of examples across various model types, purposes, and industry segments.

However, these cards and other non compulsory and non standardized formats provide only a starting point toward reliable transparency. The information they provide may be insufficient. Data cards alone may not provide sufficient transparency and clarity for data governance to be fully understood by a wide range of stakeholders. Model cards do not typically include information about the computational resources used to create the models or the labor practices involved.[884] For example, OpenAI's 100-page GPT-4 Technical Report, a format that "takes inspiration from the concepts of model cards and system cards," cites "the competitive landscape and the safety implications of large-scale models" to justify disclosing only limited information.[885] Within this context, these cards cannot be an adequate means for delivering critical information about cutting-edge models.

Moreover, these documents currently lack sufficient standardization and formalization. This lack of standardization means they can be used at cross-purposes. They can serve as a "transparent" technical report *and* a public relations tool, with the latter purpose dominating. And, they may allow for "misleading representations of model results (whether intended or unintended)."[886] As such, even the researchers who proposed model cards hold them out "as one transparency tool among many."[887]

2) Vulnerability reporting structures and post-deployment monitoring

Efficient and reliable identification of vulnerabilities related to generative AI systems is crucial for risk management. Experts advocate for the establishment of a responsible reporting framework involving government, industry, and civil society.[888] This framework is essential for addressing critical vulnerabilities and promoting the development of effective regulations. This is especially true for foundation models, which serve as the basis for specific applications and pose the risk of widespread cascading effects if a vulnerability is present.

Developers have adopted a variety of approaches to collecting and disclosing specific vulnerabilities within their system *after* they deploy an AI model. For example, Microsoft has extended its "Coordinated Vulnerability Disclosure" (CVD) policy to cover vulnerabilities in its models.[889] CVD is a long-standing

---

[884] Bommasani, et al., *The Foundation Model Transparency Index*, *see supra* note 678.

[885] *GPT-4 Technical Report supra* note 289 at 2, https://cdn.openai.com/papers/gpt-4.pdf.

[886] Mitchell, et al., *supra* note 880, at 9.

[887] *Id.*

[888] Noam Kolt et al. *Responsible Reporting for Frontier AI Developmen*t, arXiv (Apr. 3, 2024), https://arxiv.org/pdf/2404.02675. The authors propose a framework in which "[d]evelopers disclose safety-critical information to government actors and other developers, which decide on appropriate technical, organizational, and policy responses. Independent domain experts in academia and civil society receive key information and provide guidance to both developers and government actors."

[889] Microsoft, *Reporting Structure for Vulnerabilities Found after Model Release*, MICROSOFT https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/#:~:text=Reporting%20Structure%20for%20Vulnerabilities%20Found%20after%20Model%20Release (last visited June 16, 2024).

cybersecurity process by which vulnerability finders work together and share information with relevant stakeholders, such as vendors and infrastructure owners. As security researchers and ethical hackers constantly scrutinize AI systems to find weaknesses, misconfigurations, and software vulnerabilities, CVD ensures that these vulnerabilities get disclosed to the public once the developer has been able to implement a solution. The process is meant to safely disclose vulnerabilities without fear of legal action and to allow vendors to patch software vulnerabilities before they are publicly disclosed.[890] In addition to CVD, Microsoft provides "vulnerability severity classifications" that it says strengthen transparency for "customers and security researchers."[891]

OpenAI and Anthropic have published more limited policies that allow researchers to submit discovered vulnerabilities directly to the company, with the providers committing to not pursue legal action against "good faith" reporters.[892] At the time of writing, Inflection AI is in an earlier stage of calibrating its disclosure policies through a "closed pilot bug bounty program" launched in July 2023.[893]

Finally, the Frontier Model Forum (*see section 4.2.2.*)—which includes Microsoft, Anthropic, Google, and OpenAI, among others—is developing guidance on "'responsible disclosure' processes related to the discovery of vulnerabilities or dangerous capabilities within frontier models."[894]

4.1.3.C. Sourcing and authenticating content

Many harms anticipated with the use of generative AI models involve the generation of inaccurate fabricated information or the deliberate dissemination of fraudulent or deceptive content under the guise of authenticity. In response to this challenge, it is possible to enhance the reliability of the content generated by sourcing content, or to limit the credibility of the generated content by informing the user that it was produced by a generative AI tool.

1) Sourcing content: Retrieval-Augmented Generation (RAG)

Retrieval-augmented generation (RAG) is a technique that enhances the accuracy and reliability of generative AI models by incorporating factual information retrieved from external sources.[895] In other words, it is the process of optimizing the output of a generative model by referencing an authoritative knowledge base *outside* its training data sources before generating a response. For instance, the model locates relevant passages or documents containing the answer and generates a concise and coherent

---

[890] *Id.*

[891] Microsoft, *Microsoft Vulnerability Severity Classification for AI Systems*, MICROSOFT https://www.microsoft.com/en-us/msrc/aibugbar?rtc=1 (last visited June 16,, 2024).

[892] Anthropic, *Responsible Disclosure Policy*, https://www.anthropic.com/responsible-disclosure-policy (last visited June 16, 2024); OPENAI, *Coordinated vulnerability disclosure policy*, (last updated July 28, 2023), OPENAI https://openai.com/policies/coordinated-vulnerability-disclosure-policy.

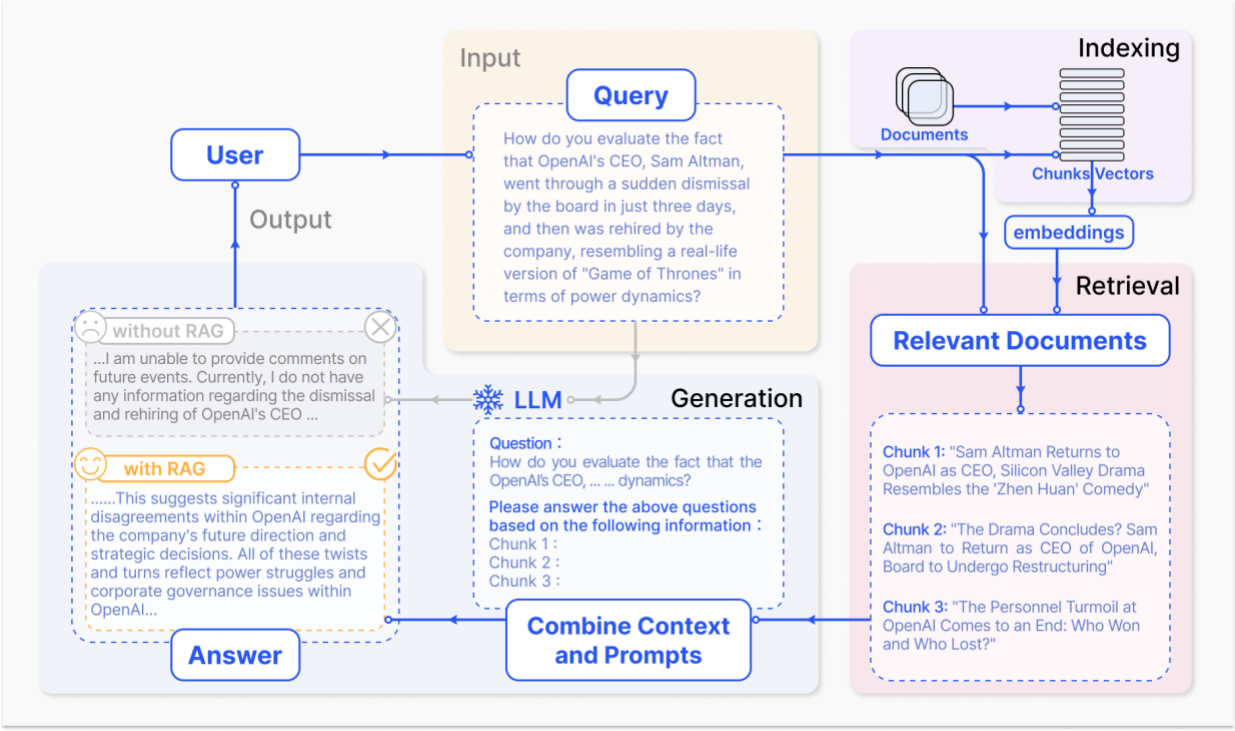
[893] Inflection AI, *Our policy on frontier safety* (Oct. 30, 2023), INFLECION AI, https://inflection.ai/frontier-safety.

[894] Microsoft, *Microsoft's AI Safety Policies*, MICROSOFT ON THE ISSUES (Oct. 26, 2023), https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/.

[895] Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, arXiv (Apr. 12, 2021), https://arxiv.org/pdf/2005.11401.

response based on that information.[896] This process ensures the model has access to the most current and reliable facts and provides users with the model's sources, allowing its claims to be verified for accuracy and ultimately trusted.

FIGURE 13. Retrieval Augmented Generation (RAG) workflow



source: Yunfan Gao et al., *Retrieval-Augmented Generation for Large Language Models: A Survey*, arXiv (Mar. 27, 2024), https://arxiv.org/pdf/2312.10997.

For example, Microsoft Copilot uses RAG to ensure that its AI assistant can provide accurate and contextually relevant responses based on the latest and most pertinent information.[897] When a user submits a query, Microsoft Copilot utilizes an information retrieval system to identify relevant information. The retrieved information is combined with the user prompt, which guides the model in generating the desired output. Then the model generates text based on the prompt and the retrieved information.

The effectiveness of this technique in addressing inaccuracies has yet to be confirmed. For instance, a recent study has assessed and reported the performance of RAG-based proprietary legal AI tools.[898] The study concludes that, while hallucinations are reduced compared to general-purpose chatbots (GPT-4), AI research tools still hallucinate "between 17% and 33% of the time."

---

[896] Yunfan Gao et al., *Retrieval-Augmented Generation for Large Language Models: A Survey*, arXiv (Dec. 18, 2023), https://arxiv.org/abs/2312.10997.

[897] Microsoft, *Retrieval Augmented Generation (RAG) in Azure AI Search* (Apr. 22, 2024), MICROSOFT https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview.

[898] Varun Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, STANFORD UNIV. (2024), https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf; *see also* Tilmann Bruckhaus, *RAG Does Not Work for Enterprises*, arXiv (May 31, 2024), https://arxiv.org/abs/2406.04369.

2) Watermarking

To prevent the dissemination of fraudulent or deceptive content presented as authentic, various approaches have been proposed to help the public identify content generated by AI models and determine which model was responsible for its creation. A first approach consists in encouraging users to reveal that the content they use or disseminate has been generated by AI. Take, for instance, OpenAI's Content Policy for its image generator, DALL·E, which suggests: "When sharing your work, we encourage you to disclose AI involvement in your work."[899] However, there is no enforceability mechanism attached to this encouragement.

Other solutions, such as having AI chatbots disclose to their users that they are AI, have important applications for mitigating harms related to influence, overreliance, and dependence but are not sufficient for addressing the broader problem of sourcing and authenticating AI-generated images, audio, videos, and text. On that front, the approach that has earned the most attention thus far from both industry and regulators is AI "watermarking," or the process of embedding a unique and detectable signal (i.e., the watermark) into AI-generated content.

a) The process of watermarking

AI watermarking is the process of embedding a recognizable and unique signal, known as a watermark, into the output of an AI model.[900] This signal serves to identify the content as AI-generated. In practice, AI watermarking creates a unique, identifiable signature that is detectable by algorithms, allowing it to be traced back to the AI model.[901] Watermarking can be as simple as adding a visible label to an image or a unique sound to generated audio, but these techniques are also relatively easy to remove or forge. Leading watermark approaches, by contrast, use sophisticated techniques to embed subtle patterns into AI-generated content that are not detectable to the human eye.[902]

In practice, watermarking involves two distinct phases: the marking phase and the identification phase.[903] The watermark is created during the model training phase by teaching the AI model to embed a specific signal or identifier in the generated content. After the AI model is deployed, specialized algorithms are used to detect the presence of the embedded watermark, thereby identifying the content as AI-generated. For example, Google DeepMind has developed a digital watermarking technology known as SynthID to distinguish AI-generated images from images created by humans. The tool can embed a digital watermark

---

[899] OpenAI, *How should I credit DALL·E in my work?*, OPENAI https://help.openai.com/en/articles/6640875-how-should-i-credit-dall-e-in-my-work.
[900] Tambiama Madiega, *Generative AI and watermarking*, EUROPEAN PARLIAMENTARY RESEARCH SERVICE (Dec. 2023), https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf.
[901] John Kirchenbauer et al., *A Watermark for Large Language Models*, arXiv (Jan. 24, 2023), https://arxiv.org/abs/2301.10226.
[902] Siddarth Srinivasan, *Detecting AI fingerprints: A guide to watermarking and beyond*, BROOKINGS (Jan. 4, 2024), https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.
[903] Madiega, *supra* note 900.

directly into AI-generated images[904] or videos[905] produced by Google's AI tools. Additionally, SynthID can scan individual images or the frames of a video to detect digital watermarking.[906] Meta's Stable Signature embeds invisible watermarks directly into AI-generated images.[907] OpenAI's DALL·E 3 incorporates both visible and invisible watermarks.[908] The visible watermark in DALL·E 3 appears as a "Content Credentials" (CR) symbol in the top left corner of the generated images, while invisible metadata is embedded to provide details about the AI tool used and the creation date. Major tech companies, including Meta, Google, and OpenAI, work together to develop common standards for watermarking. Industry standards have emerged such as C2PA (Coalition for Content Provenance and Authenticity)[909] (*see section 4.2.5.*)

Effective watermarking should allow for the detection of AI-generated content and the identification of its origin. When implemented successfully, watermarking should provide an effective method for identifying the origin of AI-generated disinformation or deepfakes. For example, in 2024, Meta has decided to implement a new policy for labeling AI-generated content on its platforms.[910] This initiative involves labeling a broad spectrum of content, including images, videos, and audio, by incorporating both visible and invisible watermarks into AI-generated materials. Invisible watermarks are embedded in the metadata of the files. Additionally, appropriate labels, such as "Made with AI," will be added as visible watermarks to inform users about the nature of the content they are viewing. Images generated using Meta's AI tools are already labeled "Imagined with AI." But the labels will also apply to images generated by AI tools developed by Google, Microsoft, OpenAI, Adobe, Midjourney, and Shutterstock—but only once these companies begin incorporating watermarks and other technical metadata into the images created by their software. Since other image generators, including open-source models, may never adopt such markers, Meta is developing tools to automatically detect AI-generated content, even in the absence of watermarks or metadata. The primary limitation of Meta's initiative is that its labeling system applies only to photos; it cannot label AI-generated audio or video because the industry has not yet standardized the inclusion of such data in these formats. For AI-generated video or audio, Meta relies on users to self-disclose when they post such content.

Beyond identifying AI-generated outputs, watermarking also provides authors with a method to trace and identify unauthorized use of their content. For example, watermarks can embed crucial information about the content creator, such as the author's name, publication date, and copyright details, directly into the digital material. This facilitates the proof of ownership and the establishment of the content's origin. The presence of a watermark can also potentially deter copyright infringers from using the content without permission, as it signals that the content is protected and traceable.

---

[904] Sven Gowal & Pushmeet Kohli, *Identifying AI-generated images with SynthID*, GOOGLE DEEP MIND (Aug. 29, 2023), https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/.

[905] Google AI Test Kitchen, *Video FX*, https://aitestkitchen.withgoogle.com/tools/video-fx (last visited on June 16, 2024).

[906] Gowal & Kohli, *supra* note 905.

[907] Matthijs Douze & Pierre Fernandez, *Stable Signature: A new method for watermarking images created by open source generative AI*, META (Oct. 6, 2023), https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/.

[908] OpenAI, *C2PA in DALL·E 3: C2PA standard, OpenAI's implementation, and C2PA metadata*, OPENAI, https://help.openai.com/en/articles/8912793-c2pa-in-dall-e-3 (last visited on June 16, 2024).

[909] COALITION FOR CONTENT PROVENANCE AND AUTHENTICITY(C2PA), https://c2pa.org/.

[910] Nick Clegg, *Labeling AI-Generated Images on Facebook, Instagram and Threads*, META (Feb. 6, 2024), https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/.

b) Limitations

State-of-the-art AI watermarking techniques have technical limitations. A recent publication concluded that neither model signatures in generated text outputs nor watermarking techniques that imprint specific patterns onto generated content are reliable in practical scenarios.[911]

In practice, AI companies encounter difficulties in creating effective watermarks, especially for generated text. Embedding a marker in text without altering its underlying meaning is challenging. A critical goal is to create methods that make digital watermarks visible to the human eye while ensuring the metadata does not interfere with the content itself. Additionally, efforts to use content analysis techniques to detect synthetic content based on its inherent properties also face technical obstacles. AI-text detectors can produce false negatives (incorrectly claiming that a human-produced text was AI-generated) or false positives (incorrectly claiming that an AI-generated text was generated by a human).

Moreover, research shows that both invisible and visible watermarks in text and audio-visual content can be manipulated, removed, or altered through various types of attacks.[912] A recent article highlights that removing a watermark from an image produced using the current C2PA watermarking standard takes approximately two seconds.[913] None of the existing solutions solve the ease with which bad actors can remove watermarks. While existing watermarks, like that of OpenAI's DALL·E,[914] Google's SynthID, and Meta's Stable Signature solutions, are more resistant to removal from images, Google admits that "high perturbation" methods, like SynthID, are not "foolproof against extreme image manipulations."[915] Creating a watermarking scheme robust enough to prevent an attacker from erasing the watermark without significantly degrading the content's quality appears to be a challenging task. Furthermore, generative AI models are vulnerable to "spoofing attacks," where adversaries generate human-produced text that is detected as AI-generated to harm the reputation or appropriate the authority of a legitimate entity.[916] Authentication methods can be easily circumvented by having a human or machine paraphrase AI-generated texts.[917]

Finally, watermarking efforts rely on users having access to the technologies required to reliably detect and authenticate watermarks. Such authentication capabilities are likely to roll out first to wealthy users in

---

[911] Vinu Sankar Sadasivan et al., *Can AI-Generated Text Be Reliably Detected?*, arXiv (June 28, 2023), http://arxiv.org/abs/2303.11156.

[912] *Researchers Tested AI Watermarks—and Broke All of Them*, WIRED (Oct. 3, 2023) https://www.wired.com/story/artificial-intelligence-watermarking-issues/.

[913] David E. Harris & Lawrence Norden, *Meta's AI Watermarking Plan Is Flimsy, at Best,* IEEE SPECTRUM (Mar. 4, 2024), https://spectrum.ieee.org/meta-ai-watermarks.

[914] Mehrdad Saberi et al., *Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks*, arXiv at 3 (Sept. 29, 2023), https://arxiv.org/pdf/2310.00076.pdf.

[915] Makena Kelly, *Watermarks Aren't the Silver Bullet for AI Misinformation*, THE VERGE (Oct. 31, 2023), https://www.theverge.com/2023/10/31/23940626/artificial-intelligence-ai-digital-watermarks-biden-executive-order.

[916] *Id.*

[917] *Id.*, at 12; Kalpesh Krishna et al., *Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense*, arXiv (Oct. 18, 2023), at 12, https://arxiv.org/pdf/2303.13408.pdf.

Western countries, since such capabilities will tend to be packaged into newer, more expensive tools. As a result, significant amounts of content produced by people with only modest resources will likely go without authentication marks, at least initially. Many systems, especially open-source systems, may lack watermarking entirely. As a result, content produced by malicious actors with even a mild degree of sophistication is likely to be missing a watermark.[918] Communities with fewer resources will be less empowered to identify malicious or deceptive content.

4.1.3.D. Removing unwanted data

While it is advisable to minimize the inclusion of web-scraped data in training datasets before pre-training a model, effective data curation alone cannot completely eliminate the risk of incorporating poor-quality data or data obtained in violation of the law. OpenAI notes in its GPT-4 System Card that it removes personal information from the training dataset "where feasible."[919] Google DeepMind, similarly, states that it made efforts to remove personal information from its pre-training data for PaLM 2.[920] Meta notes that it "made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals" from its pre-training data, but it provides no further details regarding the specific sites, the selection protocols, or the success of that effort.[921]

Unfortunately, once a model has been pre-trained, it becomes very challenging, using current techniques, to remove the specific impact of a subset of data, as the data is embedded in the model's weights. If such removal appears necessary—because, for example, it has been requested by a user or ordered by an authority—it appears to be a difficult or even impossible task, leaving only model destruction as a last resort. In this context, new techniques are emerging to delete or "disgorge" machine-learning models and algorithms trained on unwanted or inappropriate data.

To date, "model disgorgement," or "algorithmic disgorgement," has not been widely adopted as a best practice by AI companies. Instead, it has been imposed as a remedy by regulators when models have been trained on data obtained unlawfully.[922] "Model disgorgement" requires the offending company to give up

---

[918] Renée DiResta & Dave Willner, *White House AI Executive Order Takes On Complexity of Content Integrity Issues*, TECH. POL'Y PRESS (Nov. 1, 2023), https://www.techpolicy.press/white-house-ai-executive-order-takes-on-complexity-of-content-integrity-issues/.

[919] OpenAI, *GPT-4 System Card*, *supra* note 345.

[920] "We employed several data cleaning and quality filtering methods, including de-duplication, removal of sensitive-PII and filtering," Rohan Anil et al., *Palm2 Technical Report*, arXiv abs/2305.10403 (2023).The company describes a general data ingestion auditing process for pre-training and fine-tuning its "frontier" AI models in which developer teams must "submit a data ingestion request to a dedicated data team." That data team then evaluates the origin, content, and license of the data and ensures compliance with the company's AI ethics principles *See Model Evaluation and Red Teaming*, GOOGLE DEEPMIND https://deepmind.google/public-policy/ai-summit-policies/#model-evaluations-and-red-teaming:~:text=research%20and%20evaluation.-,Data%20Input%20Controls%20and%20Audit,-We%20 incorporate%20data (last visited June 15, 2024).

[921] Touvron et al., *Llama 2*, *supra* note 685.

[922] For instance, FTC Commissioner Rebecca Kelly Slaughter wrote that the premise behind algorithmic disgorgement is that "when companies collect data illegally, they should not be able to profit from either the data or any algorithm developed using it." *See* Rebecca Kelly Slaughter et al., *Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission* 23 YALE J.L. & TECH. Special Issue 1

improperly obtained data and the algorithm trained on such data,[923] on the grounds that companies who collect data illegally "should not be able to profit from either the data or any algorithm developed using it."[924] Within this framework, techniques have emerged to remove the influence of the problematic training data, effectively making the model as though that data had never been used. Faced with a "model disgorgement" order, companies may choose to retrain their models or to rely on a nascent technique called "machine unlearning."[925]

1) Retraining the model

Disgorgement through retraining involves retraining a model from scratch after first excluding undesirable data from the training dataset. It is important to emphasize that such retraining can not only be conducted to remove illegally obtained data from a dataset but also to prevent the recurrence of an unwanted model behavior. Because the capabilities of models are closely tied to their training data, this process can result in a model that does not display the unwanted behavior of the original model. The model produced after retraining can feature very different functionalities from the original model. In any case, the huge costs of training make retraining almost unfeasible in the context of the largest and most capable generative AI models.

2) Machine unlearning

"Machine unlearning"[926] aims to remove the influence of an individual data point or a collection of data points from a model, rather than eliminating the data and retraining the model. In other words, the model, after unlearning, should perform as if it had never learned the problematic data in the first place. Originally, unlearning was proposed as a method to protect privacy and copyright by neutralizing the influence of undesirable training data. However, "machine unlearning" can also help remove undesirable capabilities from generative AI systems.[927] For instance, it can eliminate capabilities that could enable malicious users to create bioweapons and conduct cyberattacks.

---

https://yjolt.org/sites/default/files/23_yale_j.l._tech._special_issue_1.pdf. In its 2019 settlement with Cambridge Analytica, the FTC ordered that "any algorithms or equations that originated, in whole or in part, from" data that had been illegally collected from Facebook users had to be deleted. *See* Bruce D. Solker et al., *Algorithmic Disgorgement: An Increasingly Important Part of the FTC's Remedial Arsenal — AI: The Washington Report*, NATIONAL LAW REVIEW, (January 24, 2024), https://natlawreview.com/article/algorithmic-disgorgement-increasingly-important-part-ftcs-remedial-arsenal-ai.

[923] Brandon LaLonde, *Explaining model disgorgement*, IAPP (Dec. 13, 2023) https://iapp.org/news/a/explaining-model-disgorgement/; *see also* Joshua A. Goland, *Algorithmic Disgorgement: Destruction of Artificial Intelligence Models as the FTC's Newest Enforcement Tool for Bad Data* (Mar. 1, 2023). RICHMOND J. OF LAW AND TECH., Vol. XXIX, Issue 2 (2023), Available at SSRN: https://ssrn.com/abstract=4382254.

[924] Slaughter et al. *supra* note 922.

[925] LaLonde, *supra* note 923.

[926] Thanh Tam Nguyen et al., *A Survey of Machine Unlearning*, arXiv (Oc. 21, 2022), https://arxiv.org/pdf/2209.02299; Varun Gupta, *Adaptive Machine Unlearning*, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 34 (NeurIPS 2021). https://proceedings.neurips.cc/paper/2021/hash/87f7ee4fdb57bdfd52179947211b7ebb-Abstract.html; Lucas Bourtoule et al., *Machine Unlearning*, arXiv (Dec. 15, 2020) https://arxiv.org/pdf/1912.03817.

[927] Bengio et al., *International Scientific Report supra* note 7 at 75.

Machine unlearning involves sophisticated algorithms that can selectively reduce the impact of unwanted data. However, unlearning methods often fail to perform robustly and may introduce unwanted side effects on the desirable data of the model. Current scholarship is focused on developing algorithms that can achieve unlearning efficiently, without compromising the model's accuracy or necessitating a complete retraining. Moreover, research in this area explores the scalability of unlearning methods to handle large and complex models for which the interdependencies between data points and model parameters are highly intricate.

Machine unlearning is still a nascent field with many open questions, including the optimal balance between unlearning efficiency and model performance. In June 2023, Google announced the first "Machine Unlearning Challenge," an open competition designed to advance the understanding of challenges facing the field of machine learning, standardize evaluation metrics for different unlearning algorithms, and foster novel solutions.[928]

4.2. Collective initiatives

The practices discussed in the previous section originate from individual initiatives by generative AI companies, which often collaborate on these practices, as exemplified in the case of reinforcement learning and watermarking. Within this framework, certain practices and methodologies have gradually emerged or are in the process of becoming recognized as best practices. Simultaneously, more structured initiatives, such as industry alliances, are being undertaken to address specific issues and risks. This section outlines the primary collective initiatives in this domain.[929]

These initiatives, in their efforts to establish standards for the industry, represent a move toward self-regulation.[930] However, they do not result in the creation of *binding* principles or standards for companies. Their primary benefit lies in *facilitating* the emergence of common, interoperable standards, making best practices more visible, and providing input for regulators thinking about drafting legal frameworks.

4.2.1. The Partnership on AI

The "Partnership on AI to Benefit People and Society" (PAI) is an independent, nonprofit organization founded in 2016 by a coalition of technology companies, civil society organizations, and academic institutions. Initially, it received multi-year grants from founding members Apple, Amazon, Meta,

[928] Fabian Pedregosa and Eleni Triantafillou, *Announcing the first Machine Unlearning Challenge*, GOOGLE RESEARCH (June 29, 2023) https://blog.research.google/2023/06/announcing-first-machine-unlearning.html.
[929] Cat Zakrzewski & Nitasha Tiku, *AI Companies Form New Safety Body, While Congress Plays Catch-up*, WASH. POST (July 26, 2023), https://www.washingtonpost.com/technology/2023/07/26/ai-regulation-created-google-openai-microsoft/.
[930] Alyssa Wong, *Regulatory gaps and democratic oversight: On AI and self-regulation*, SCHWARTZ REISMAN INST. FOR TECH. & SOC'Y. (Sept. 21, 2023), https://srinstitute.utoronto.ca/news/tech-self-regulation-democratic-oversight.

Google/DeepMind, IBM, and Microsoft.[931] While founded by industry leaders, the organization also convenes a community of academics, civil rights groups, and major media organizations to research AI best practices. By its own definition, it is a "resource to policymakers—for example, to conduct research that informs AI best practices and to explore the societal consequences of certain AI systems, as well as the policies surrounding the development and use of these systems."[932] The PAI holds member status within the Civil Society Information Society Advisory Council (CSISAC), which is part of the OECD's Network of Experts.[933] It is also a member of the U.S. AI Safety Institute Consortium (*see section 5.3.2.B.3.f.*).[934] It held observer status within the Council of Europe's Committee on Artificial Intelligence during the drafting of the international treaty on AI (*see section 6.8.*).

Access Now, a nonprofit focusing on digital civil rights, withdrew from the PAI a little over a year after joining.[935] The organization, which supports human rights impact assessments and a ban on facial recognition technologies, did not agree with the PAI's "ethics, risk-based, or sandboxing approach" nor with its openness to the use of facial recognition technologies.[936]

The PAI publishes regular reports, articles, and guidelines on the development and use of generative AI.[937] Its more notable publications include the following:

- In February 2023, the PAI published guidelines in the "Responsible Practices for Synthetic Media,"[938] which provide recommendations for developers and deployers of AI-generated media. The guidelines emphasize many of the same values that exist in other AI ethical guidelines, including transparency, accuracy, nondiscrimination, privacy, and accountability. At the same time, the PAI guidelines identify stakeholders that may meaningfully reduce harm at the different stages of a synthetic media's development. The document has fostered consensus among important industry members and key organizations, with Google, Meta, Microsoft, OpenAI, Stanford HAI, and others having each supported the framework.[939]

---

[931] *Partnership on AI*, https://www.partnershiponai.org (last visited Feb. 14, 2024).

[932] *Game Changers: Artificial Intelligence Part III – AI and Public Policy, Hearing Before the Subcomm. on Info. Tech. of the H.R. Oversight and Gov't Reform Comm.*, 115th Congress 2 (2018) (statement of Terah Lyons Executive Director, Partnership on AI), https://oversight.house.gov/wp-content/uploads/2018/04/Lyons-PAI-Statement-AI-III-4-18.pdf.

[933] *Public Policy*, P'SHIP ON AI, https://partnershiponai.org/program/policy/ (last visited May 19, 2024).

[934] *AISIC Members*, US AI SAFETY INST., https://www.nist.gov/aisi/aisic-members (last visited May 19, 2024).

[935] *Access Now resigns from the Partnership on AI*, ACCESS NOW (Oct. 13, 2023), https://www.accessnow.org/press-release/access-now-resignation-partnership-on-ai/.

[936] *Id.*

[937] Sarah Villeneuve et al., *Eyes Off My Data: Exploring Differentially Private Federated Statistics to Support Algorithmic Bias Assessments Across Demographic Groups*, P'SHIP ON AI (last visited March 8, 2024), https://partnershiponai.org/wpcontent/uploads/dlm_uploads/2023/12/PAI_whitepaper_eyes-off-my-data-1.pdf.

[938] *PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action*, P'SHIP ON AI , https://syntheticmedia.partnershiponai.org (last visited Feb. 14, 2023).

[939] Including Adobe, BBC, CBC/Radio-Canada, Bumble, TikTok, WITNESS, and Synthesia.

- In October 2023, the PAI published *Guidance for Safe Foundation Model Deployment*.[940] This document is designed to assist model providers in responsibly developing and deploying various AI models. The *Guidance* is unique among similar frameworks in that its emphasis is on a tailored approach. It suggests practices for ongoing reassessment as AI capabilities evolve, accommodating a variety of AI models and deployment scenarios, including frontier and open access models. It enumerates 22 guidelines, at varying stages of a model's development, for different foundation models and release types (i.e., closed development, research release, restricted release, and open access). It provides starting points to address a wide range of safety risks, such as potential harms from bias, overreliance on AI systems, worker treatment, and malicious activities by bad actors. Notably, the *Guidance* advocates for staged releases and restricted access for frontier models until adequate safeguards are demonstrated.[941] The guidelines have been commended for their collaborative effort[942] and were recently released in their final form. Thus far, Google, Meta, Microsoft, Apple, OpenAI, the Alan Turing Institute, and the Ada Lovelace Institute, among others, have expressed their support for the guidelines.

- For model developers, PAI published a study[943] in December 2023 that looks at the use of differentially private federated statistics, a privacy-preserving, analytical technique that combines the methods of differential privacy (*see section 4.1.1.D.*) and federated statistics.[944] The study assesses the advantages of differentially private federated statistics, drawing on the case study of Apple when it applied the technique to allow users to safely upload their IDs to Apple Wallets. At the same time, it examines the socio-technical risks and the implications for algorithmic fairness. Overall, the study recommends obtaining consent for the data used and verifying the reliability of data that has been processed to ensure that personal identifiers of respondents cannot be linked to their responses. It shows, however, that it is difficult to ensure that the processing of a database does not miscategorize a part of the population, especially when using differentially private federated statistics.

---

[940] *PAI's Guidance for Safe Foundation Model Deployment*, P'SHIP ON AI , https:// partnershiponai.org/wp-content/uploads/1923/10/PAI-Model-Deployment-Guidance.pdf?ref=maginative.com (last visited May 19, 2024).

[941] *Id.*

[942] *Partnership on AI Releases Guidance for Safe Foundation Model Deployment, Takes the Lead to Drive Positive Outcomes and Help Inform AI Governance Ahead of AI Safety Summit in UK*, P'SHIP ON AI (Oct. 24, 2023), https://partnershiponai.org/pai-model-deployment-guidance-press-release/.

[943] Sarah Villeneuve et al., *Eyes Off My Data: Exploring Differentially Private Federated Statistics to Support Algorithmic Bias Assessments Across Demographic Groups,* P'SHIP ON AI (Dec. 13, 2023), https://partnershiponai.org/paper/eyes-off-my-data/.

[944] Federated statistics is a "machine learning (ML) technique that enables organizations to access and use data from multiple, discrete devices without the need to collect and store this data in a centralized database." *Differentially Private Federated Statistics*, P'SHIP ON AI, https://partnershiponai.org/paper_page/differentially-private-federated-statistics/ (last visited May 19, 2024).

- In May 2024, the Partnership on AI introduced guidelines for prioritizing equity in algorithmic systems.[945] These guidelines aim to help AI practitioners navigate the complexities of demographic data collection to advance equitable systems while avoiding harm to marginalized groups.

Overall, PAI serves as an especially valuable forum for allowing various stakeholders to delve into issues related to AI safety and ethics and to develop common guidelines. However, its recommendations do not carry any particular authority.

4.2.2. Frontier Model Forum

Google, OpenAI, Microsoft, and Anthropic announced on July 26, 2023, that they were forming the "Frontier Model Forum" to ensure safe and responsible development of frontier AI models.

The organization defines "frontier models" as "large-scale machine-learning models that exceed the capabilities currently present in the most advanced existing models, and can perform a wide variety of tasks."[946] The goal of the Frontier Model Forum is "to help (i) advance AI safety research to promote responsible development of frontier models and minimize potential risks, (ii) identify safety best practices for frontier models, (iii) share knowledge with policymakers, academics, civil society, and others to advance responsible AI development; and (iv) support efforts to leverage AI to address society's biggest challenges."[947] According to members of the Forum, this work also includes collaboration to develop guidance on "responsible disclosure" of vulnerabilities or dangerous capabilities within frontier models.[948]

The Forum members and their philanthropic partners announced three months later that they were establishing an "AI Safety Fund" to support independent research into AI safety. The Forum described the Fund as "an important part of fulfilling" the Voluntary AI Commitments agreement signed at the White House earlier that month.[949] The Fund, with more than $10 million in initial funding, focuses on advancing research into "new model evaluations and techniques for red-teaming AI models."[950] On April 1, 2024, the

---

[945] Eliza McCullough, *Prioritizing Equity in Algorithmic Systems through Inclusive Data Guidelines* (May 14, 2024), https://partnershiponai.org/prioritizing-equity-in-algorithmic-systems-through-inclusive-data-guidelines/.
[946] OpenAI, *Frontier Model Forum* (July 26, 2023), *Adaptive Machine Unlearning*, OPEN AI, https://openai.com/index/frontier-model-forum.
[947] *Microsoft, Anthropic, Google, and OpenAI launch Frontier Model Forum*, MICROSOFT (July 26, 2023), https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/.
[948] *Microsoft's AI Safety Policies*, MICROSOFT ON THE ISSUES (Oct. 26, 2023), https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/.
[949] *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI*, WHITE HOUSE (Sept. 12, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.
[950] *Anthropic, Google, Microsoft and OpenAI announce Executive Director of the Frontier Model Forum and over $10 million for a new AI Safety Fund*, GOOGLE (Oct. 25, 2023), https://blog.google/outreach-initiatives/public-policy/google-microsoft-anthropic-open-ai-frontier-model-forum-executive-director/.

Forum announced that it has awarded the first round of research grants from the AI Safety Fund.[951] The Forum has also released issue briefs on measuring training compute[952] and red teaming.[953]

4.2.3. The AI Alliance

Meta and IBM, in collaboration with more than 50 founding members, launched the AI Alliance in December 2023.[954] The AI Alliance describes itself as an international group of "leading organizations across industry, startup, academic, research, and government coming together to support open innovation and open science in AI." In April 2024, the AI Alliance announced the addition of more than 20 new members, integrating a diverse mix of academic institutions, startups, enterprises, and scientific organizations from around the world.[955]

To support its mission, the AI Alliance has established two initial member-driven working groups: "AI Safety and Trust Tooling" and "AI Policy Advocacy." These working groups bring together researchers, developers, policymakers, and industry experts to collaboratively and transparently address the challenges of generative AI and promote its widespread benefits. The AI Safety and Trust Tooling working group will provide objective information and best practice guidance on AI safety and establish comprehensive benchmarking capabilities for testing AI models and applications. The AI Policy Advocacy working group will create public forums, publish and disseminate information and opinions, and represent the broader AI ecosystem's reliance on open source and open innovation to policymakers.

The AI Alliance notably reflects Meta's position within the industry as an outspoken advocate of open-source models. Importantly, the AI Alliance states its objectives as "plans to start or enhance projects" that "[r]esponsibly advance the ecosystem of open foundation models" and "encourage open development of AI in safe and beneficial ways," among other objectives related to model evaluation and benchmarking.[956]

4.2.4. MLCommons

MLCommons is an AI engineering consortium of more than 125 members and affiliates from the industry, academia, and nonprofit sectors.[957] It is dedicated to accelerating machine learning (ML) innovation and

---

[951] *AI Safety Fund initiates first round of research grants*, FRONTIER MODEL F. (Apr. 1, 2024), https://www.frontiermodelforum.org/updates/ai-safety-fund-initiates-first-round-of-research-grants/.
[952] *Issue Brief: Measuring Training Compute*, FRONTIER MODEL F. (May 2, 2024), https://www.frontiermodelforum.org/updates/issue-brief-measuring-training-compute/.
[953] *Issue Brief: What is red teaming?*, FRONTIER MODEL F. (Oct. 27, 2023), https://www.frontiermodelforum.org/updates/red-teaming/.
[954] *AI Alliance Launches as an International Community of Leading Technology Developers, Researchers, & Adopters Collaborating Together to Advance Open, Safe, Responsible AI*, IBM (Dec. 5, 2023), https://newsroom.ibm.com/AI-Alliance-Launches-as-an-International-Community-of-Leading-Technology-Developers,-Researchers,-and-Adopters-Collaborating-Together-to-Advance-Open,-Safe,-Responsible-AI; AI Alliance, *Building the Open Future of AI*, AI ALLIANCE, https://thealliance.ai/ (last visited June 20, 2024).
[955] *Id. Building the Open Future of AI*, AI ALLIANCE.
[956] *Id.*.
[957] MLCOMMONS, https://mlcommons.org/ (last visited on July 20, 2024).

promoting collaboration within the ML community. Founded in 2020, MLCommons brings together a diverse group of stakeholders, including researchers, academics, engineers, and industry leaders, to develop benchmarks, best practices, datasets, and open-source tools to advance the field of machine learning.

MLCommons is best known for its MLPerf benchmark suite, which provides standardized performance benchmarks for evaluating the speed and efficiency of machine-learning hardware and software. MLCommons also focuses on creating and maintaining open datasets and tools that support machine-learning research and development. MLCommons also works on establishing best practices and standards for machine-learning development and deployment.

In this context, the MLCommons consortium is playing a leading role in the development of benchmarks for generative AI systems. In October 2023, it announced the formation of an AI Safety Working Group focused on developing safety benchmarks for LLMs that build on the framework—Holistic Evaluation of Language Models (HELM)—developed by Stanford University's CRFM.[958] Initial industry participation in the working group includes Anthropic, Coactive AI, Google, Inflection, Intel, Meta, Microsoft, Nvidia, OpenAI, and Qualcomm Technologies, Inc. In April 2024, the working group announced a safety benchmark proof of concept, with plans for a "comprehensive v.10 release" later in 2024.[959]

4.2.5 Coalition for Content Provenance and Authenticity

In 2021, Microsoft co-founded the Coalition for Content Provenance and Authenticity (C2PA) with Adobe, Arm, BBC, Intel, and Truepic. This initiative was formed by combining two previously existing initiatives—the Adobe-led Content Authenticity Initiative (CAI) and the Microsoft- and BBC-led Project Origin.[960] Both these initiatives focused on tackling the issue of content provenance through different means. The Coalition aims to develop the C2PA technical specification to authenticate content provenance. Since its formation, many companies have joined the C2PA steering committee with the intention of adopting the technical standards developed by the C2PA. These companies include X, Sony,[961] OpenAI,[962] and Google.[963]

Content provenance refers to the use of metadata to present detailed information to an end user about the origins of, and alterations to, a piece of digital content.[964] It helps in understanding the "provenance" of content, such as an image or video, and whether such content has been altered. The provenance of content

---

[958] *MLCommons announces the formation of AI safety working group*, MLCOMMONS (Oct. 26, 2023), https://mlcommons.org/2023/10/mlcommons-announces-the-formation-of-ai-safety-working-group/.

[959] *Creating a benchmark suite for safer AI*, MLCOMMONS, https://mlcommons.org/ai-safety/ (last visited May 19, 2024).

[960] COAL. FOR CONTENT PROVENANCE & AUTHENTICITY, https://c2pa.org/.

[961] *News*, COAL. FOR CONTENT PROVENANCE & AUTHENTICITY , https://c2pa.org/post/ (last visited May 19, 2024).

[962] *Understanding the source of what we see and hear online*, OPENAI (May 7, 2024), https://openai.com/index/understanding-the-source-of-what-we-see-and-hear-online/.

[963] *News*, COAL. FOR CONTENT PROVENANCE & AUTHENTICITY , *see supra* note 961.

[964] *Generative AI, content provenance & a public service internet*, THE ROYAL SOC'Y-BBC (July 11, 2023), https://royalsociety.org/-/media/policy/projects/digital-content-provenance/digital-content-provenance_workshop-note_.pdf.

may include information such as who created it and how, and when and where it was created or edited.[965] The rationale behind using content provenance is to help users assess the authenticity of the content and determine whether it is trustworthy.

Specifically, C2PA develops a technical standard called Content Credentials,[966] a tamper-proof metadata standard designed to verify the creation and modification history of digital content. This standard helps users identify whether a particular image or video is AI-generated.[967] The C2PA standard was recently implemented by Meta Platforms on its social media platforms and by OpenAI with its DALL·E 3 image generator (*see section 4.1.3.C.2.*). Google is exploring ways to integrate the C2PA's Content Credentials into its products and services, such as Gemini.

4.2.6. Other initiatives

There are also initiatives that are not necessarily formalized through alliances or coalitions, but are nevertheless industry-driven efforts to address various challenges arising from generative AI. Some of these examples are:

- **Societal Resilience Fund**: This is a $2 million fund announced by Microsoft and OpenAI to further AI education and literacy among voters and vulnerable communities.[968] This fund is initially being used to provide grants to organizations working on AI education and tackling deceptive content generated using AI. Among others, the Partnership on AI and the C2PA are recipients of grants from this fund.[969]

- **Commitments to promote child safety**: In collaboration with two nonprofit organizations—Thorn and All Tech Is Human—major AI companies publicly committed to enacting "Safety by Design" principles to prevent the creation and spread of AI-generated child sexual abuse material and other sexual harms against children.[970] These companies include Anthropic, OpenAI, StabilityAI, Microsoft, Amazon, Google and Meta, among others.[971] These principles are given in a larger report released by Thorn and All Tech Is Human on how safety by design can be used to tackle child sexual abuse-related harms from generative AI.[972]

---

[965] *FAQ*, COAL. FOR CONTENT PROVENANCE & AUTHENTICITY, *see supra* note 961.

[966] *Content Credentials : C2PA Technical Specification*, COAL. FOR CONTENT PROVENANCE & AUTHENTICITY , https://c2pa.org/specifications/specifications/2.0/specs/C2PA_Specification.html (last visited May 19, 2024).

[967] Tate Ryan-Mosley, *Cryptography may offer a solution to the massive AI-labeling problem*, MIT TECH. REV. (July 28, 2023), https://www.technologyreview.com/2023/07/28/1076843/cryptography-ai-labeling-problem-c2pa-provenance/. .

[968] *Microsoft and OpenAI launch Societal Resilience Fund*, MICROSOFT (May 7, 2024), https://blogs.microsoft.com/on-the-issues/2024/05/07/societal-resilience-fund-open-ai/. .

[969] *Id.*

[970] *Thorn and All Tech Is Human Forge Generative AI Principles with AI Leaders to Enact Strong Child Safety Commitments*, THORN (Apr. 23, 2024), https://www.thorn.org/blog/generative-ai-principles/.

[971] *Id.*

[972] *Safety by Design for Generative AI: Preventing Child Sexual Abuse*, THORN, ALL TECH IS HUMAN (Apr. 2024), https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf.

Another industry group known as the "Tech Coalition" announced that it would fund new research on generative AI and online child sexual exploitation and abuse. Tech Coalition members include Adobe, Amazon, Bumble, Google, Meta, Microsoft, OpenAI, Roblox, Snap Inc., and TikTok.[973]

- **AI Elections Accord**: Also known as "A Tech Accord to Combat Deceptive Use of AI in 2024 Elections," the AI Elections Accord refers to a set of eight industry commitments to tackle harmful AI-generated content that tries to deceive voters.[974] These accords were announced in February 2024 to address concerns about the misuse of AI-generated content to target voters participating in the national elections of over 40 countries this year.[975] There are 20 signatories to this accord, including Google, Meta, Microsoft, IBM, X, Anthropic, OpenAI, and StabilityAI.[976]

KEY TAKEAWAYS

- The growing public attention and evolving risks associated with generative AI have prompted AI companies to develop practices aimed at mitigating those risks. AI companies typically justify their risk mitigation measures in public documents that outline their guiding principles and policies for training and deploying AI models. Although the specific practices adopted by companies can vary significantly, safety practices can be examined by distinguishing those applied in the pre-deployment, deployment, and post-deployment stages.

- At the pre-deployment stage, a crucial mitigating practice is the implementation of good data governance, particularly data curation. This process ensures the quality and suitability of training data by evaluating the reliability of data sources, deciding which data to include or exclude, removing specific data from an aggregated dataset, and creating or augmenting data to address gaps, imbalances, or other limitations. Developers may also use "differential privacy" to address privacy concerns. This technique involves adding a degree of "statistical noise" to the training data to ensure that the model's output remains nearly identical, regardless of whether a single individual's data is included or excluded.

- It is also essential to identify the model's potential vulnerabilities. Benchmarking, which measures and compares the performance of different models using standardized datasets, metrics, and tasks, provides an objective assessment of an AI model's capabilities and limitations. Red teaming is also crucial; this process involves adversarial engagement with AI systems to expose their limitations

[973] *Tech Coalition Announces New Generative AI Research*, TECH COAL. (May 9, 2024), https://www.technologycoalition.org/newsroom/tech-coalition-announces-new-generative-ai-research.
[974] *A Tech Accord to Combat Deceptive Use of AI in 2024 Elections*, AI ELECTIONS ACCORD (Feb. 16, 2024), https://www.aielectionsaccord.com/uploads/2024/02/A-Tech-Accord-to-Combat-Deceptive-Use-of-AI-in-2024-Elections.FINAL_.pdf.
[975] AI ELECTIONS ACCORD, https://www.aielectionsaccord.com/.
[976] *Technology industry to combat deceptive use of AI in 2024 elections*, AI ELECTIONS ACCORD (Feb. 16, 2024), https://www.aielectionsaccord.com/uploads/2024/02/Press-Release-AI-Elections-Accord-16-Feb-2024.pdf.

and vulnerabilities. However, despite the industry's promotion of red teaming as a vital tool for identifying risks in generative AI systems, there are limited concrete details on how red teaming is being implemented and what is its actual effectiveness.

- The behavior and outcomes of an AI model must align as closely as possible with the goals and values established by its developers. This alignment can be achieved by improving fine-tuning practices, particularly through the use of reinforcement learning to align the model with desired objectives. An emerging application of reinforcement learning through AI feedback is "Constitutional AI." In Constitutional AI, an AI assistant is trained to evaluate outputs using a set of predefined rules, or "constitutional" principles. The model's behavior is, therefore, assessed without the necessity of human intervention to identify harmful outputs.

- Once the model is pre-trained, tested, and fine-tuned, it is up to its developers to decide the appropriate time for its release. The AI developer and research community has created frameworks that systematize the decision-making process regarding how, when, for whom, and whether models should be released. The primary purpose of these "Responsible Scaling Policies (RSPs)" is to establish the correct course of action when certain risks are identified before release. Some risks warrant specific deployment measures, while others may require withholding deployment of an excessively risky model. Currently, some major generative AI companies have adopted such frameworks, and open-source developers are beginning to disclose their release protocols. However, critics argue that existing policies are insufficient and call for more robust commitments and clearly defined risk thresholds.

- Once their models are released, AI providers encounter significant challenges in preventing the misuse of their technology for illegal or harmful purposes. To address this, their Terms of Use and Usage Policies typically outline acceptable and unacceptable uses of AI products and services. Providers can also guide users in crafting prompts that avoid generating illegal or policy-violating outputs. Moreover, AI systems can be designed to prevent objectionable actions. For example, they can refuse to follow certain instructions, decline to generate specific content, and customize responses to user prompts to ensure compliance with established guidelines.

- Enhanced transparency regarding AI models is an effective way to mitigate technical vulnerabilities and reduce the risk of AI misuse. Providing detailed information to regulators, users, and the general public about potential risks can foster a better understanding of AI models. Model, system, and data cards contribute to this understanding by providing technical details and other insights into how the model works, its limitations, and vulnerabilities. Additionally, some developers proactively collect and disclose specific vulnerabilities after deploying an AI model. Finally, addressing the risks of generating inaccurate or fabricated information, as well as the deliberate spread of fraudulent or deceptive content presented as authentic, can be managed by informing users that the content was AI-generated. "Watermarking," the process of embedding a unique and detectable signal into AI-generated content, provides an effective method for identifying the origin of AI-generated disinformation or deepfakes. Additionally, Retrieval-Augmented Generation (RAG) enhances the accuracy and reliability of generative AI systems by incorporating factual information retrieved from external sources. This process involves optimizing the output of a

generative model by referencing an authoritative knowledge base outside its training data sources before generating a response.

- AI companies have organized themselves into industry groups. Among these alliances, the Partnership on AI to Benefit People and Society (PAI) serves as a resource for policymakers, while the AI Alliance is an international group of leading organizations supporting open innovation and open science in AI. The Frontier Model Forum aims to advance AI safety research, identify best practices for frontier models, and share knowledge with stakeholders. The Coalition for Content Provenance and Authenticity focuses on developing technical specifications to authenticate content provenance. These initiatives represent a step toward self-regulation. However, they typically do not establish precise or binding principles or standards for companies. Their main advantage lies in promoting the development of common, interoperable standards, highlighting best practices, and providing valuable input for regulators who are considering the creation of binding legal frameworks.

- Overall, the individual or collective initiatives adopted within the industry highlight the vigorous efforts of leading AI organizations to develop practices aimed at minimizing risks and fostering the emergence of common safety standards. Most of the risks mentioned in Chapter 3 appear to be considered, as illustrated in the table below (*see figure 14*). However, these practices do not necessarily exemplify self-regulation. For some practices, such as red teaming or reinforcement learning, the focus is on technological advancement and enhancing the quality and, consequently, the safety of AI models. Moreover, while developing these practices involves extensive communication and collaboration among AI companies and sometimes leads to the establishment of common standards, they remain entirely voluntary commitments without any independent oversight. Consequently, there is no reliable way to ensure these practices are actually implemented or consistently enforced. The general public must rely on AI companies to voluntarily adhere to their own commitments and announcements.

- Although these emerging standards and practices do not specifically exemplify self-regulation, they can contribute to the development of self-regulatory instruments. They are widely discussed and collaboratively refined within the AI community, often evolving into recognized best practices. Coalitions support this process by promoting transparency and cooperation among companies. Consequently, these practices and initiatives may eventually be acknowledged by regulators, either as part of nonbinding frameworks, like the NIST frameworks, or within formal legal frameworks, such as the EU's AI Act. On the other hand, these initiatives enable the industry, especially its dominant players, to position themselves as key interlocutors with regulators. Overall, these efforts function as both tools of influence and platforms for developing solutions that benefit the general public.

FIGURE 14. Possible risks and industry practices

The table below, similar to this chapter, does not aim to offer an exhaustive overview of all AI company practices, as these practices are highly diverse. Rather, the modest objective of this table is to summarize

the principal techniques discussed in this chapter, illustrating how these techniques can effectively mitigate the risks and challenges addressed in Chapter 3.

Possible risks and challenges of generative AI	**Industry practices**
Technical vulnerabilities (section 3.1.1)	• Curating datasets (section 4.1.1.A.) • Benchmarking (section 4.1.1.B.1.) • Red teaming (section 4.1.1.B.2.) • Model alignment and Reinforcement Learning (section 4.1.1.C.) • Self-destructing models (section 4.1.3.A.3.) • Vulnerability reporting and post-deployment monitoring (section 4.1.3.B.2.) • Retraining the model (section 4.1.3.D.1.)
Factually incorrect content (section 3.1.2.)	• Curating datasets (section 4.1.1.A.) • Model Alignment and Reinforcement Learning (section 4.1.1.C.) • Retrieval-augmented generation (4.1.3.C.1.)
Opacity (section 3.1.3)	• Model Cards, Data Cards, System Cards, and Technical Reports (section 4.1.3.B.1.)
Malicious use and abuse (section 3.2.1)	• Red teaming (section 4.1.1.B.2.) • Model Alignment and Reinforcement Learning (section 4.1.1.C.) • Usage policies, terms of service, and licenses (section 4.1.3.A.2.) • Product interaction design (section 4.1.3.A.1.) • Self-destructing models (section 4.1.3.A.3.)
Misinformation and disinformation (section 3.2.2)	• Curating datasets (section 4.1.1.A.) • Model Alignment (section 4.1.1.C.) • Usage policies, terms of service, and licenses (4.1.3.A.2.) • Usage monitoring (section 4.1.3.A.1.) • Retrieval-augmented generation (RAG) (section 4.1.3.C.1.) • Watermarking (section 4.1.3.C.2.)
Bias and discrimination (section 3.2.3)	• Curating datasets (section 4.1.1.A.) • Model Alignment (section 4.1.1.C.)
Influence, overreliance, and dependence (section 3.2.4)	• Watermarking (section 4.1.3.C.2.)

New capabilities (section 3.2.5)	• Responsible scaling policy (section 4.1.2.A.)
Possible risks of open-source models (section 3.2.6.A)	• Open source responsible scaling policy (section 4.1.2.B.)
Possible risks of Highly Capable Models (Section 3.2.6.B)	• Responsible scaling policy (section 4.1.2.A.) • Recommendation of the the PAI Model Deployment guidance in favor of staged releases and restricted access for frontier models until adequate safeguards are demonstrated (section 4.2.1.)
Privacy and data protection (section 3.3.1)	• Curating datasets (section 4.1.1.A.) • Differential privacy (section 4.1.1.D.) • Retraining the model (section 4.1.3.D.1.) • Machine unlearning (section 4.1.3.D.2.)
Copyrights (section 3.3.2)	• Curating datasets (section 4.1.1.A.) • Web crawlers equipped to recognize and exclude protected data (ex: GPTBot) • Watermarking (section 4.1.3.C.2.) • Retraining the model (section 4.1.3.D.) • Machine unlearning (section 4.1.3.D.)

Chapter 5 Regulatory initiatives

Some regions and countries have chosen to implement robust legal frameworks to regulate artificial intelligence, acknowledging the swift technological advancements and associated risks. These regulatory strategies vary: Some, such as the European Union framework, explicitly adopt a risk-based methodology, whereas others, including the Chinese framework, primarily follow a principle-based approach. Overall, the various approaches mentioned in the introduction (*see section 1.2*) —self-regulation, co-regulation, and regulation— are all reflected in the strategies employed by different countries. However, they differ by more pronounced tendencies toward a specific approach: While the United States favors self-regulation, Europe combines regulation and co-regulation, and China tends to adopt a top-down regulatory approach.

This chapter focuses on AI governance and regulatory initiatives, excluding other aspects of AI strategies employed by different countries, particularly those related to investment. The following overview of regulatory frameworks for generative AI encompasses both established laws and pending bills. Significant attention is dedicated to the European Union (section 5.1), which has recently enacted the AI Act. This section also offers a comprehensive analysis of Chinese laws. (section 5.2) and AI policies in the United States (section 5.3). Additionally, the discussion extends to various countries (section 5.4) where efforts to govern and regulate AI are currently active.

5.1. The European Union

To date, the European Union is one of the few regions in the world to have adopted a comprehensive regulatory framework specifically for AI. Despite the challenges inherent in formulating pertinent rules for a swiftly advancing technology with rapidly proliferating applications, the EU chose to legislate without delay. This approach was motivated by the desire to establish legal principles and standards governing the deployment and utilization of AI, rather than deferring to industry entities to dictate their preferences. This ambitious objective proved challenging, as evidenced by the multiple versions of the Regulation proposed during negotiations over the Artificial Intelligence Act.[977]

The recently enacted AI Act is not the sole regulatory framework governing AI within the European Union. At the EU level, a multitude of laws exists, encompassing both overarching and industry-specific provisions regulating the activities of technology companies, including AI developers. Furthermore, there are national laws adopted by individual Member States of the EU. This section discusses certain European laws and regulations already in effect, such as the GDPR and the DSA. However, it does not examine all other statutes in the EU. For instance, the Digital Markets Act,[978] regulates entities known as "gatekeepers," targeting companies such as Alphabet, Amazon, Meta, and Microsoft, which are actively involved in developing and deploying generative AI models and systems. Although this Regulation applies to certain

[977] Council Regulation 2024/1689 of June 13, 2024, (**Artificial Intelligence Act**), 2024 O.J. (L 12.7.2024), http://data.europa.eu/eli/reg/2024/1689/oj (*see below section 5.1.2)*.
[978] Council Regulation 2022/1925 of Sept. 14, 2022, (**Digital Markets Act**), 2022 O.J. (L 265, 12.10.2022), http://data.europa.eu/eli/reg/2022/1925/oj.

AI companies, it will not be specifically addressed here, as its primary purpose is not to regulate AI providers. The Data Act,[979] which seeks to foster a competitive data market by mandating that data holders share data collected through connected products, virtual assistants, or related services, will not be examined either. And, the Data Governance Act,[980] which regulates the regime applicable to public sector data and the activities of data intermediary services, will also not be studied.

In addition to the AI Act, the EU has recently adopted the Revised Directive on Product Liability,[981] extending its scope to include AI software. A new directive on AI liability[982] has also been proposed.[983] Furthermore, the European Cyber Resilience Act[984] was adopted alongside the AI Act and the new Defective Products Directive.

5.1.1. Existing legal frameworks in the EU

While numerous laws are applicable to AI and its various applications within the EU, certain ones warrant particular attention.

5.1.1.A. The General Data Protection Regulation (GDPR)

Companies developing artificial intelligence and offering generative AI models and systems in the EU must ensure their activities comply with the EU's General Data Protection Regulation (GDPR).[985] The GDPR

[979] Council Regulation 2023/2854 of Dec. 13, 2023, (**Data Act**), 2023 O.J. (L 22.12.2023), http://data.europa.eu/eli/reg/2023/2854/oj.

[980] Council Regulation 2022/868 of May 30, 2022, (**Data Governance Act**), 2022 O.J. (L 152/1), http://data.europa.eu/eli/reg/2022/868/oj.

[981] European Parliament legislative resolution P9_TA(2024)0132 of Mar. 12, 2024 on the proposal for a directive on liability for defective products (COM(2022)0495 – C9-0322/2022 – 2022/0302(COD)), (**New Product Liability Directive**), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0132_EN.html *(see section 5.1.3.A)*

[982] Proposal for a Directive of the European Parliament and of the Council on Adapting Non-contractual Civil Liability Rules to Artificial Intelligence, (**AI Liability Directive**), COM (2022) 496 final (Sept. 28, 2022), https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52022PC0496 *(see section 5.1.3.B).*

[983] The two principal forms of legislation under EU law are regulations and directives. Regulations are directly applicable and binding in their entirety in Member States (i.e., countries that are members of the European Union). Directives set out certain positions that must be implemented by EU Member States but leave the form and method of implementing those positions up to the Member States to determine.

[984] European Parliament legislative resolution P9_TA(2024)0130 of Mar. 12, 2024 on the proposal for a regulation on horizontal cybersecurity requirements for products with digital elements (**Cyber Resilience Act)**, (COM(2022)0454 – C9-0308/2022 – 2022/0272(COD), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0130_EN.pdf (*see section 5.1.4*).

[985] Council Regulation 2016/679 of Apr. 27, 2016 (**General Data Protection Regulation**) O.J. (L 119/1), http://data.europa.eu/eli/reg/2016/679/oj (*see section 5.1.1.A*).

applies when personal data is processed by a controller[986] or processor[987] for an establishment in the European Union.[988] And it applies regardless of whether or not the data processing takes place in the EU. In fact, the GDPR can still apply even if a company is not established in the EU: According to Article 3 of the GDPR, the regulation applies to any company that offers goods or services to EU data subjects or monitors their behavior, insofar as that behavior occurs within the EU.

The GDPR's one-stop-shop mechanism is designed to enable a company processing European individuals' data to deal with a single lead supervisory authority, which is the privacy watchdog situated in the EU Member State where the company has its primary establishment.[989] In principle, privacy regulators located in other EU jurisdictions redirect complaints to the lead supervisory authority of the country where the company's main establishment is located.[990] In the absence of any company office in the EU, as was the case with OpenAI until recently,[991] the one-stop-shop mechanism does not apply and all national privacy regulators may have jurisdiction. Within this framework, European data protection authorities play a pivotal role in the regulation and oversight of generative AI systems.

For example, in March 2023, following a data breach, the Italian Data Protection authority (also known as Garante per la Protezione dei Dati Personali —or simply Garante) started an investigation and identified that OpenAI was not adhering to its obligations under the General Data Protection Regulation. On March 30, 2023, the Garante adopted a temporary decision against OpenAI.[992] The Garante's order[993] required OpenAI to immediately and temporarily stop processing the personal data of Italy-based users, pending further investigation. OpenAI responded by restricting access in Italy to its chatbot. One month later, after OpenAI adopted new measures,[994] the Garante confirmed that it could resume operations and process the

---

[986] The data controller determines the *purposes* for which and the *means* by which personal data are processed. *See: What is a data controller or a data processor?*, EUROPEAN COMMISSION, https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/controllerprocessor/what-data-controller-or-data-processor_en.

[987] The data processor processes personal data only on behalf of the controller and is usually a third party, external to the company.

[988] GDPR, art. 3.

[989] *Id.* art. 60.

[990] In practice, this does not always happen. Furthermore, other GDPR regulators maintain the authority to intervene locally if they identify imminent risks. *See*, Joe Jones, *Practical considerations from EU enforcement: One-stop shop*, IAPP (Feb 2023), https://iapp.org/resources/art./practical-considerations-eu-enforcement-pt2/ (accessed June 29, 2024); *see also* Edward Machin, *This week in data/cyber/tech: The GDPR's one stop shop just became a lot harder, AI in recruitment, and settling data subject complaint*s, ROPES & GRAY (Feb. 23, 2024), https://www.ropesgray.com/en/insights/viewpoints/102j0vm/this-week-in-data-cyber-tech-the-gdprs-one-stop-shop-just-became-a-lot-harder.

[991] Paul Sawers, *OpenAI to open its first EU office as it readies for regulatory hurdles*, TECHCRUNCH (September 14, 2023), https://techcrunch.com/2023/09/14/openai-dublin-eu-regulation/; Natasha Lomas, *OpenAI moves to shrink regulatory risk in EU around data privacy*, TECHCRUNCH (Jan. 2, 2024),https://techcrunch.com/2024/01/02/openai-dublin-data-controller/.

[992] *Provvedimento del 30 marzo 2023 [9870832]*, GARANTE PER LA PROTEZIONE DEI DATI PERSONALI (Mar. 20, 2023), https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870832.

[993] *Id.*

[994] *Preparedness*, OPENAI, https://openai.com/safety/preparedness (accessed June 20. 2024).

data of Italy-based users.[995] A similar incident occurred with the AI chatbot Replika, which was barred from processing the personal data of Italy-based users for several months due to violations of various GDPR principles.[996]

The Garante is pursuing its investigations. On January 29, 2024, it announced that it notified OpenAI it was in violation of data protection law and activated sanction proceedings.[997] On March 8, 2024, Garante initiated an investigation into OpenAI's latest AI model, Sora, which has the capability to generate realistic and imaginative scenes from brief textual prompts.[998]

Data protection authorities in other EU Member States are closely monitoring the release of AI models in the EU and the compliance of AI companies with the GDPR. The Irish Data Protection Commission (DPC) required Google to postpone the June 2023 launch of its AI platform Bard (now called Gemini) in the EU because Google had not submitted sufficient information to the DPC.[999] Poland's data protection authority initiated an investigation following a complaint that OpenAI's ChatGPT fabricated information about an individual and refused to correct the inaccuracies.[1000] In other countries, such as Germany[1001] and France,[1002]

---

[995] *ChatGPT: OpenAI Reopens the Platform in Italy Ensuring More Transparency and More Rights to European Users and Non-users*, GPDP (Apr. 28, 2023), https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9881490.

[996] Natasha Lomas, *Replika, a 'virtual friendship' AI chatbot, hit with data ban in Italy over child safety,* TECHCRUNCH (Feb. 3, 2023), https://techcrunch.com/2023/02/03/replika-italy-data-processing-ban/.

[997] *ChatGPT: Italian DPA notifies breaches of privacy law to OpenAI*, GPDP (Jan. 29, 2024), https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9978020#english.

[998] *Intelligenza artificiale, il Garante privacy avvia istruttoria su "Sora" di OpenAI. Chieste alla società informazioni su algoritmo che crea brevi video da poche righe di testo*, GARANTE PER LA PROTEZIONE DEI DATI PERSONALI (Mar. 30, 2024), https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9991867#english.

[999] *Google Bard released in the EU after privacy concerns were addressed*, DATENSCHUTZ-NOTIZEN, https://www.datenschutz-notizen.de/google-bard-released-in-the-eu-after-privacy-concerns-were-addressed-3643725/ (last accessed June 29, 2024).

[1000] Natasha Lomas, *Poland opens privacy probe of ChatGPT following GDPR complaint*, TECHCRUNCH (Sept. 21, 2023), https://techcrunch.com/2023/09/21/poland-chatgpt-gdpr-complaint-probe/.

[1001] For instance, the Data Protection Authority of Baden-Württemberg (Landesbeauftragte für den Datenschutz und die Informationsfreiheit Baden-Württemberg) pointed out that a complete understanding of OpenAI's adherence to data protection laws depends on identifying the specific purposes of the data processing and the datasets used to train the model. It raised concerns regarding the possibility that prompts could reveal details about an individual, potentially including insights into their political, religious, ideological, or scientific beliefs, or information about their family and personal life. *See LfDI informs himself at OpenAI how ChatGPT works under data protection law*, LFDI (Apr. 24, 2023), https://www.baden-wuerttemberg.datenschutz.de/lfdi-informiert-sich-bei-openai-wie-chatgpt-datenschutzrechtlich-funktioniert/. *See also AP asks for clarification about ChatGPT*, AUTORITEIT PERSOONSGEGEVENS (June 7, 2023), https://autoriteitpersoonsgegevens.nl/actueel/ap-vraagt-om-opheldering-over-chatgpt.

[1002] Natasha Lomas, *France's privacy watchdog eyes protection against data scraping in AI action plan*, TECHCRUNCH (May 17, 2023), https://techcrunch.com/2023/05/17/cnil-ai-action-plan/.

data protection authorities have also raised concerns about generative AI companies' compliance with GDPR.[1003]

Authorities are sometimes petitioned by nongovernmental organizations (NGOs) or individuals. On April 29, 2024, the Austrian nongovernmental organization NOYB ("None Of Your Business") filed[1004] a data protection complaint[1005] with Austria's data protection authority. The complaint accused OpenAI of violating the GDPR. On June 6, 2024,[1006] NOYB filed multiple privacy complaints against Meta, seeking an urgent decision from data protection authorities in Austria, Belgium, France, Germany, Greece, Italy, Ireland, the Netherlands, Norway, Poland, and Spain. The aim was to prevent Meta from using user posts on Facebook, Instagram, and other Meta platforms to train its generative AI models. As a result, the Irish DPC engaged with Meta on the issue, leading Meta to pause its plans to use public content shared by adults on Facebook and Instagram to train its large language model.[1007] In July 2024, Meta announced that it would not release its multimodal Llama model in the EU, citing "the unpredictable nature of the European regulatory environment."[1008] Although Meta did not provide specific details, it is likely that these concerns are related to data protection compliance concerns.

In this context, the European Data Protection Board (EDPB) created a dedicated task force aimed at enhancing collaboration and facilitating the exchange of information concerning potential enforcement actions undertaken by data protection authorities.[1009] The "ChatGPT Taskforce" released a preliminary report in May 2024.[1010] In the meantime, the European Data Protection Supervisor adopted guidelines on the use of generative AI by EU institutions and bodies.[1011] The following paragraphs analyze the main GDPR compliance issues related to generative AI models and systems.

[1003] Autoriteit Persoonsgegevens, *Blogpost: zorgen om generatieve AI* (Dec. 7, 2023), https://www.autoriteitpersoonsgegevens.nl/actueel/blogpost-zorgen-om-generatieve-ai.

[1004] NOYB - European Center for Digital Rights, *ChatGPT provides false information about people, and OpenAI can't correct it,* NOYB (Apr. 24, 2024) https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it.

[1005] NOYB - European Center for Digital Rights, *OpenAI complaint*, NOYB (Apr. 29, 2024), https://noyb.eu/sites/default/files/2024-04/OpenAI%20Complaint_EN_redacted.pdf.

[1006] NOYB - European Center for Digital Rights, *noyb urges 11 DPAs to immediately stop Meta's abuse of personal data for AI*, NOYB (June 6, 2024), https://noyb.eu/en/noyb-urges-11-dpas-immediately-stop-metas-abuse-personal-data-ai.

[1007] Meta asserts that it is following the example set by other companies, such as Google and OpenAI, which have already utilized data from European users to train their AI models. *See* Stefano Fratta, *Building AI Technology for Europeans in a Transparent and Responsible Way*, META (June 10, 2024), https://about.fb.com/news/2024/06/building-ai-technology-for-europeans-in-a-transparent-and-responsible-way/.

[1008] Jess Weatherbed, *Meta says European regulators are ruining its AI bot,* THE VERGE, (July 18, 2024), https://www.theverge.com/2024/7/18/24201041/meta-multimodal-llama-ai-model-launch-eu-regulations

[1009] European Data Protection Board (EDPB), *EDPB resolves dispute on transfers by Meta and creates task force on ChatGPT*, EDPB (Apr. 13, 2023), https://edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt_en.

[1010] European Data Protection Board (EDPB), *Report of the work undertaken by the ChatGPT Taskforce* (May 23, 2024), EDPB, https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf.

[1011] EDPS, *Data protection and artificial intelligence* (AI), (Oct. 24, 2023), https://www.edps.europa.eu/system/files/2023-10/2023-10-24-edps-at-work-data-protection-and-artificial-intelligence_en.pdf.

1) Lawfulness of processing (Article 6 of the GDPR)

A core principle of the GDPR holds that every processing of personal data must rest on some established legal basis.[1012] In the case of generative AI models, this requirement applies to both pre-training and fine-tuning with any information likely to identify a living individual, directly or indirectly. Identifying information includes details like names, birthdates, email addresses, phone numbers, home addresses, location data, images, any kind of content that says something about a given individual, or even characteristics specific to a person's physical, mental, genetic, economic, cultural, or social identity.

In its preliminary report,[1013] the EDPB underscored the importance of differentiating between the various stages of personal data processing in the case of chatbots, such as ChatGPT. According to the EDPB, these stages include:

- the collection of training data (such as through web scraping or reusing existing datasets),
- the pre-processing of data (including filtering),
- the training phase,
- the generation of prompts and chatbot outputs, and
- the training of the chatbot using prompts.

Within this context, a generative AI company can establish legal basis for processing personal data in one of several ways, including by *obtaining the consent* of the individual whose data is being processed;[1014] establishing the necessity of processing the data for the *performance of a contract* with the individual or to take steps preparatory to such a contract;[1015] or demonstrating that the processing is necessary for the purposes of the *legitimate interest* pursued by the developer.[1016] These three possible legal bases will be considered one after the other.

- Considering the large number of individuals whose data might be utilized in the training of generative AI models, securing the express consent from each can prove challenging. Moreover, personal consent must be obtained for every use or activity, which is particularly difficult in the context of the possible future development of unpredictable downstream applications. However, chat interface providers can collect consent from users when they register for the service. In this case, consent may work as a legal basis to process the inputs containing personal data entered by users.

- If consent cannot be obtained, AI developers may seek to demonstrate a "legitimate interest" for using personal data. They must establish that "the processing is necessary for the purposes of the legitimate interests pursued by the data controller" (i.e., the developer), as provided by Article 6(1)(f) of the GDPR. However, such a legal basis cannot be taken for granted. The GDPR provides

[1012] GDPR, art. 6.
[1013] EDPB, *Report of the work undertaken by the ChatGPT Taskforce*, see *supra* note 1010, §14.
[1014] GDPR, art. 6(1)(a).
[1015] *Id.* art. 6(1)(b).
[1016] *Id.* art. 6(1)(f).

that this legal basis will be disregarded "where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child."[1017] Each situation must be evaluated based on its specific circumstances, "taking into consideration the reasonable expectations of data subjects based on their relationship with the controller," as outlined by Recital 47 of the GDPR. Regulatory authorities will probably require AI companies to provide serious arguments to support the claim that their legitimate interests are properly balanced against the rights and freedoms of individuals.

- Developers may also argue that the processing is necessary for the performance of a contract or to take steps preparatory to such a contract.[1018]

In the case of OpenAI, the Garante required the company to change the legal basis on which it sought permission to process the personal data of Italy-based users for the purpose of algorithmic training. OpenAI was asked to remove any reference to contracts and to rely on individual consent or legitimate interest as its legal bases. In the most recent version of OpenAI's documentation, the company relies on a claim of legitimate interest.[1019] Concerning Google's Gemini, the "Gemini Apps Privacy Hub"[1020] explains that it relies on the legal grounds of contract performance and legitimate interests but indicates that it might seek consent requests for future features.

However, the acceptability of using "legitimate interest" as a legal basis by developers of generative AI remains unconfirmed. On July 4, 2023, the Court of Justice of the European Union (CJEU)[1021] ruled that "legitimate interest" was an inappropriate basis for Meta to rely on to justify tracking and profiling individuals in order to target its behavioral-based advertising business, despite the fact that the services of Facebook are free of charge. Recently, a guidance issued by the Dutch Data Protection Authority stated that data scraping by private companies and individuals will almost always be in violation of the GDPR.[1022] However, the European Commission (EC) weighed in, saying it believes "commercial interests can be

---

[1017] *Id.* art. 6(1)(f).

[1018] *Id.* art. 6(1)(b).

[1019] *Privacy Policy*, OPENAI (Nov. 14, 2023), https://openai.com/policies/privacy-policy (The company explains that one of its reasons for processing personal information is, "Our legitimate interests in protecting our Services from abuse, fraud, or security risks, or in developing, improving, or promoting our Services, including when we train our models. This may include the processing of Account Information, Content, Social Information, and Technical Information.")

[1020] *What Are Google's Legal Bases of Processing Gemini Apps Data Under European Union (EU) or United Kingdom (UK) Data Protection Law?*, GEMINI APPS HELP, https://support.google.com/gemini/answer/13594961?hl=en#legal_basis&zippy=%2Cwhat-are-googles-legal-bases-of-processing-gemini-apps-data-under-european-union-eu-or-united-kingdom-uk-data-protection-law (last visited Feb. 23, 2024).

[1021] Court of Justice of the European Union (CJUE), case C-252/21, Meta Platforms Inc. et al. v. Bundeskartellamt, ECLI:EU:C:2023:537 (July 4, 2023), https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:62021CJ0252.

[1022] Autoriteit Persoonsgegevens (AP), *Handreiking scraping door particulieren en private organisaties en particulieren*, (May 1, 2024) https://www.autoriteitpersoonsgegevens.nl/uploads/2024-05/Handreiking%20scraping%20door%20particulieren%20en%20private%20organisaties.pdf.

regarded as 'legitimate' interests when (subject to a concrete balancing) they are not overridden by the fundamental rights and freedoms of the data subject."[1023]

In its recent preliminary report,[1024] the EDPB does not rule out the "legitimate interest" basis but stresses the need for data controllers to implement adequate safeguards to positively influence the balancing test. These safeguards may include technical measures, setting precise criteria for data collection, and ensuring the exclusion of certain data categories or sources, such as public social media profiles. Additionally, measures should be in place to delete or anonymize personal data collected via web scraping before the training stage. Finally, regarding user inputs, such as "prompts," the EDPB emphasized that data subjects should be clearly and demonstrably informed that their content may be used for training purposes.[1025] This factor is essential to consider in the context of the balancing of interests under Article 6(1)(f) GDPR.

2) Principles relating to the processing of personal data (Article 5 of the GDPR)

Article 5 of the GDPR sets out seven key principles related to the processing of personal data: (i) lawfulness, fairness, and transparency; (ii) purpose limitation; (iii) data minimization; (iv) accuracy; (v) storage limitation; (vi) integrity and confidentiality; and (vii) accountability. Specifically, the GDPR provides that personal data must be "processed lawfully, fairly and in a transparent manner in relation to the data subject."[1026] The processing must also be **"**adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed."[1027] Compliance with these principles is particularly difficult for generative AI models, primarily because foundation models have an undefined range of potential purposes.

Regarding the principle of fairness, the EDPB insists that personal data should not be processed in a manner that is unjustifiably detrimental, unlawfully discriminatory, unexpected, or misleading to the data subject.[1028] Additionally, the report emphasizes that there should be no risk transfer, meaning the responsibility for ensuring GDPR compliance should not be placed on data subjects. For instance, the Terms and Conditions should not state that data subjects are responsible for their chat inputs, especially when these inputs include personal information.[1029]

Data must be "accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate (…) are erased or rectified without delay."[1030] In its preliminary report, the EDPB notes that "the purpose of the data processing is to train ChatGPT" rather than "to provide factually accurate information."[1031] Due to the probabilistic nature of the system, "the current training

[1023] European Commission, *Letter to the Autoriteit Persoonsgegreven*, (Mar., 6, 2020), https://static.nrc.nl/2022/pdf/letter-dutch-dpa-legitimate-interest.pdf.
[1024] EDPB, see *supra* note 1010 at §16-17.
[1025] *Id.* at §22.
[1026] GDPR, art. 5 (1).
[1027] GDPR, art. 5(1)(c).
[1028] EDPB, *see supra* note 1010 at §23.
[1029] *Id.* at §24.
[1030] GDPR, art. 5(1)(d).
[1031] EDPB, *see supra* note 1010 at §30.

approach leads to a model which may also produce biased or made up outputs." Therefore, the EDPB insists that developers and providers should offer "proper information on the probabilistic output creation mechanisms and on their limited level of reliability," including "explicit reference to the fact that the generated text, although syntactically correct, may be biased or made up."[1032] However, the EDPB also states that "the principle of data accuracy must be complied with."[1033] This accuracy requirement was underlined by the Garante in the case of OpenAI. The Italian authority highlighted that outputs containing hallucinations or inaccuracies may present a challenge in this context. OpenAI now provides an option for users to request the deletion of any information deemed inaccurate, noting, however, that it is currently technically infeasible to correct such inaccuracies.

The complaint[1034] filed by NOYB against OpenAI argues that "as long as ChatGPT keeps showing inaccurate data,"[1035] OpenAI violates Article 5(1)(d) GDPR. In this case, when asked to provide the complainant's date of birth, ChatGPT kept displaying inaccurate information. OpenAI responded that the only way to prevent the inaccurate information from appearing would be to block any information concerning the complainant. This, the complaint says, would violate freedom of expression and the general public's right to be informed, as the complainant is a public figure.[1036] Even though OpenAI has filters enabling it to block displaying personal data, it is not possible to block the complainant's date of birth without affecting other pieces of information that ChatGPT displays about this public figure, which explains why OpenAI did not take action.[1037]

3) Sensitive data (Article 9 of the GDPR)

The GDPR expressly prohibits the processing of sensitive data, unless one of the exceptions specifically provided for in Article 9(2) applies. Sensitive data are defined as "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership," and "genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation."[1038]

Despite the prohibition, it is possible that sensitive data can be used to train or run generative AI models in certain situations. Lawful processing of sensitive personal data is permitted when, among other circumstances, 1) the concerned individual gives explicit consent, 2) the processing relates to data which are manifestly made public by the person concerned, or 3) the processing is carried out during the legitimate activities (with appropriate safeguards) of a nonprofit organization. The very limited number of

[1032] *Id.* at §31.
[1033] *Id.* at §30.
[1034] NOYB, *OpenAI complaint, see supra* note 1005.
[1035] NOYB, *OpenAI complaint, Id.* at §31.
[1036] *Id.* at §28.
[1037] *Id.* at §8.
[1038] Article 9(1) of the GDPR prohibits the "processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation."

circumstances in which sensitive data processing is permitted is likely to make things extremely difficult for generative AI companies. This is why some scholars advocate for the creation of a novel exemption to the prohibition outlined in Article 9 of the GDPR.[1039]

In its report, the EDPB suggests that developers take measures to filter out sensitive data whose processing is prohibited under GDPR. This should take place during data collection (by setting criteria for what data is collected) and immediately afterward (by deleting any inappropriate data).[1040]

4) Transparency (Article 13-14 of the GDPR)

Articles 13 and 14 of the GDPR outline the information that must be provided to data subjects. Article 13 embodies the principle of transparency, outlining the data controller's obligation to provide clear and comprehensive information to individuals about the processing of their personal data in situations where personal data are collected *directly* from the data subjects (i.e., direct collection). In cases where data are collected directly from users who have registered with a chatbot service, complying with Article 13 does not seem insurmountable.

The situation becomes more complex when data is obtained indirectly, such as through web scraping. Article 14 of the GDPR requires data controllers to provide specific information to the data subjects shortly after obtaining the data from a third party. In the context of large models processing very large volumes of data often collected by data aggregators, such as the German nonprofit LAION,[1041] it does not seem possible to inform each data subject individually. Fortunately, Article 14(5)(b) provides an exemption when "the provision of such information proves impossible or would involve a disproportionate effort." Recital 62 states that "the number of data subjects, the age of the data and any appropriate safeguards adopted should be taken into consideration." It is very likely that generative AI developers will attempt to rely on this exemption. However, they will have to be as transparent as possible. Article 5(b) of the GDPR specifies that "the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available."

The Garante asked OpenAI to launch an awareness campaign in broadcast and online media to inform users that personal data may have been used and to explain how such data could be deleted via an online tool. OpenAI responded in two ways:

- Information Notice:[1042] OpenAI posted a notice on its website for both users and non-users of ChatGPT. This notice details the types of personal data processed for training the AI system and how this processing occurs. It explicitly informs individuals of their right to opt out of such processing.[1043]

[1039] Claudio Novelli et al., *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity* 9 Centre For Digital Ethics Working Paper (Feb. 19, 2024), https://ssrn.com/abstract=4694565.

[1040] EDPB, *see supra* note 1010 at §19.

[1041] LAION releases large training datasets, such as LAION 5B, which contains 5.8 billion image-text pairs and is the input dataset for Stable Diffusion's model.

[1042] *Data Controls*, https://help.openai.com/en/collections/8471418-data-controls (last visited June 20, 2024).

[1043] *Id.*

- Privacy Policy Accessibility: The privacy policy for ChatGPT users was made readily available on both the sign-up page for new users and the welcome page for users based in Italy.[1044]

In its preliminary report,[1045] EDPB emphasizes the importance of informing data subjects that their user inputs may be used for training purposes when personal data is collected through direct interaction with ChatGPT. For personal data collected via web scraping from publicly accessible sources such as websites, the EDPB suggests that the exemption under Article 14(5)(b) of the GDPR may apply, provided all conditions of this provision are met.

5) Right of access by data subjects (Article 15 of the GDPR)

Under Article 15 of the GDPR, individuals have the right to obtain a copy of their personal data that are subject to processing (i.e., any form of use) by controllers, along with other pertinent information. The complaint[1046] filed by NOYB against OpenAI states that a data subject filed an access request with OpenAI, which provided them with general information related to the data processed within the context of their user account. However, the complaint says, OpenAI provided no information about the personal data of the data subject that was processed "through the ChatGPT large language model" (§22). Therefore, the complaint alleges that OpenAI violated Article 15(1) to (3) and Article 12(3) of the GDPR.[1047]

6) Managing individuals' requests to rectify or erase data (Article 16-17 of the GDPR)

GDPR Article 16, titled "Right to rectification," introduces the right to rectify inaccurate data and the right to complete incomplete data. Article 17, titled "Right to erasure," confers upon the data subject the right to have their personal data erased. First, individuals can require the erasure of personal data without undue delay. Second, data controllers have the obligation to erase personal data in certain circumstances, such as when "the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed," where "the personal data have been unlawfully processed," or where there is no legal ground for the processing.

However, after training or fine-tuning a model with personal data, removing that data from the model is impossible. If an individual requests the erasure of their data, the appropriate solution would involve the destruction of the existing model and the training of a new model using a dataset from which the personal data of the individual in question have been excluded. This presupposes that the personal data in question are easily identifiable, which is not always the case with huge datasets. Moreover, "the deletion of data from a training dataset represents a superficial solution, as it does not necessarily obliterate the potential for data retrieval or the extraction of associated information encapsulated within the model's

[1044] *Europe Privacy Policy*, OPENAI (Dec. 15 2023, effective Feb. 15, 2024), https://openai.com/policies/privacy-policy.

[1045] EDPB, *see supra* 1010, at §27-28.

[1046] NOYB, *Complaint against OpenAI*, *see supra* note 1005.

[1047] These articles stipulate that, when requested, the controller shall provide information without undue delay and in any event within one month of receipt of the request.

parameters."[1048] The initial dataset used for training, or any data associated with the removed information, might inadvertently reveal itself or "leak," thereby compromising the effectiveness of the deletion process and perpetuating potential breaches of privacy.

In April 2023, OpenAI addressed some of these concerns by making it easier for users of ChatGPT to prevent their data from being used to train and improve models. Users can opt out their data by disabling "chat history." When disabled, says OpenAI, the chat history still retains user conversations for 30 days but only for the purpose of monitoring for abuse. After 30 days, the data is permanently deleted.[1049] OpenAI still retains data from inputs of non-business users to its browser interfaces, when they have not disabled their "chat history."[1050] Google allows users to delete their Gemini usage history and to set how long their data is stored, from three to 36 months.[1051]

It is certainly preferable for developers to ensure from the outset that no personal data (i.e., of a nature to identify an individual directly or indirectly) are present in their training data, using anonymization techniques, if necessary. However, the reliability of AI companies' commitments to data deletion and non-use is questionable, especially given past incidents in which large tech companies failed to fulfill promises regarding the handling and deletion of their users' data.[1052] OpenAI does not claim to perfectly prevent the inclusion of personal data in training. Rather it says it tries to "reduce the amount of personal information in [its] training datasets before they are used to improve and train [its] models."[1053] The opacity surrounding the training data of these models further complicates the issue, making it unclear how an individual or enterprise customer could ascertain whether their data are used in violation of applicable policies.

7) Protection of minors (Article 8 of the GDPR)

[1048] Novelli et al., *supra* note 1039, at 11.

[1049] *New Ways to Manage Your Data in ChatGPT*, OpenAI (Apr. 25, 2023), https://openai.com/blog/new-ways-to-manage-your-data-in-chatgpt. "As of March 1st, 2023, we retain customer API data for 30 days but no longer use customer data sent via the API to improve our models." OpenAI, *How Can I Use the Chat Completion (ChatGPT) API?*, OpenAI, https://help.openai.com/en/articles/7232945-how-can-i-use-the-chatgpt-api (last visited Feb. 23, 2024)

[1050] *Data Controls FAQ*, OpenAI, https://help.openai.com/en/articles/7730893-data-controls-faq (last visited Feb. 23, 2024) ("If I disable history, does the setting apply to all my conversations, or can I choose specific conversations to enable it for?").

[1051] *Manage & Delete Your Gemini Apps Activity*, Google Gemini Apps Help, https://support.google.com/gemini/answer/13278892?hl=en&co=GENIE.Platform%3DAndroid (last visited Feb. 23, 2024).

[1052] Geoffrey A. Fowler, *Google Promised to Delete Sensitive Data. It Logged My Abortion Clinic Visit.*, Wash. Post (May 9, 2023 11:23 AM), https://www.washingtonpost.com/technology/2023/05/09/google-privacy-abortion-data/; *FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook*, FTC (July 24, 2019), https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions-facebook.

[1053] *How your data is used to improve model performance*, OpenAI, https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance (last visited Feb. 23, 2024).

Under Article 8(2) of the GDPR, the data controller must undertake reasonable efforts to verify that children's consent is "given or authorized by the holder of parental responsibility over the child, taking into consideration available technology."

On February 2, 2023, the Italian Data Protection Authority (Garante) issued an urgent order blocking the AI-powered chatbot Replika from processing the personal data of Italian users.[1054] The order was based on concerns that Replika posed risks to minors and vulnerable individuals, Specifically, the Garante found that Replika, designed for individuals above 13, lacked sufficient mechanisms to verify users' ages, merely requiring users to provide their names, email addresses, and genders. Tests showed that even when Replika received explicit statements indicating a user was a minor, it did not block interactions, potentially exposing minors to inappropriate content, including sex-related material. The Garante also highlighted that Replika fails to disclose essential information about its processing of personal data, particularly children's data, violating GDPR transparency requirements.

In the case of OpenAI, the Garante noted the absence of any age verification procedures for ChatGPT users, even though OpenAI's Terms of Service professes to restrict the use of ChatGPT for children under 13 years old. In response, OpenAI introduced an age verification system for Italy-based users.[1055] The service sign-up area now includes a birthdate request, denying access to those under 13 and requiring confirmation of parental consent for users between 13 and 18.

8) Automated decision-making (Article 22 of the GDPR)

Article 22 of the GDPR provides that automated decision-making is prohibited for decisions producing "legal effects concerning [the user] or similarly significantly affects [the user]." On December 7, 2023, the Court of Justice of the European Union (CJEU) issued a ruling expanding the scope of the prohibition significantly.[1056] The court held that the creation of a credit score is an automated decision for the purposes of Article 22, even if the bank is the ultimate decision-maker, because the score plays a "determining role" in the decision process. In this context, generative AI models, like any other AI system, cannot be used to adopt automated decisions of this kind, as long as their outputs can be considered as playing a determining role in the decision process. The rule will, therefore, have repercussions whenever a generative AI model is used to evaluate individuals, for example, in higher education, recruitment, or financial credibility. It will be necessary to ensure that the output generated by the model does not play a decisive role in the decision's adoption, unless one of the exceptions in Article 22 is established.

There are exceptions to the prohibition, such as in cases where the concerned individual gives explicit consent. There are also exceptions where a specific statute authorizes automated decision-making or where

---

[1054] GPDP (*Garante*), *Intelligenza artificiale, dal Garante privacy stop al chatbot "Replika." Troppi i rischi per i minori e le persone emotivamente fragili* (Feb. 3, 2023), *https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9852506#english*;
Natasha Lomas, *Replika, a 'virtual friendship' AI chatbot, hit with data ban in Italy over child safety,* TechCrunch (Feb. 3, 2023) https://techcrunch.com/2023/02/03/replika-italy-data-processing-ban/.

[1055] *Terms of Use of Europe*, OpenAI (Feb. 15, 2024), https://openai.com/it/policies/eu-terms-of-use.

[1056] Court of Justice of the European Union (CJUE), case C-634/21, OQ v. Land Hessen, ECLI:EU:C:2023:957 (Dec. 7, 2023), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62021CJ0634.

automated decision-making is necessary to enter into or perform a contract. It is not necessarily easy to obtain the consent of a concerned individual or to establish the necessity of the process for contractual purposes. And even if automated decision-making is allowed, data subjects have the right to contest the decision and obtain human intervention in the decision. Additionally, Article 15(1)(h) of the GDPR explicitly states that, when automated decision-making (including profiling) is used, the right of access for data subjects includes access to meaningful information about the logic, significance, and envisaged consequences of that processing for the data subject. Here again, such explanations may be difficult to provide, given the complexity and opacity of generative AI models.

5.1.1.B Copyright and patent issues

The development and operation of generative AI models can impact intellectual and industrial property rights in various ways. First, the data utilized for training and fine-tuning models may be subject to copyright protection. Second, the outputs from generative AI systems might closely resemble existing content, especially if such content is included in the training data, potentially leading to copyright infringement. Last, the legal status of these outputs, particularly the degree to which they are eligible for copyright or patent protection, is still unresolved.

1) EU copyright law related to training

Training an AI system with scraped data often leads to issues of intellectual property rights infringement, particularly when the training process uses copyrighted works without authorization. In the EU, the primary recourse to mitigate this problem is to invoke the "Text and Data Mining exception," despite its inherent limitations.

a) The determination of protected content

The EU Copyright Directive 2001/29[1057] requires EU Member States to provide authors, performers, phonograph producers, film producers, and broadcasting organizations with an "exclusive right to authorize or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part" of their works, performances, films, etc.[1058] Of course, the precise rules governing copyright depend on how the Copyright Directive is implemented into, and interpreted under, the applicable national laws of the Member States, as well as any other relevant rules under national law. The fact remains that, on principle, authors have the discretion to permit or deny the use of their work.

In addition, the Database Directive 96/9/EC[1059] requires EU Member States to protect the compilation of information comprising a database. "Database" for these purposes means a collection of independent works, data, or other materials arranged in a systematic or methodical way and individually accessible by electronic

[1057] Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (**Copyright Directive**), O.J. (L 167, 22.6.2001), http://data.europa.eu/eli/dir/2001/29/oj.
[1058] Copyright Directive, art. 2.
[1059] Directive (EU) 96/9/EC of 11 March 1996 on the legal protection of databases (**Database Directive**) O.J. (L 77, 27.3.1996), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009.

or other means. However, for databases to qualify for copyright protection, they must constitute "the author's own intellectual creation," which means they must exhibit personal influence and entail creative choices.[1060] Creativity may involve the organization of the database, the types of columns it includes, or its indexing methods.

Databases that merely present objective results and factual information and, thus, lack originality, are ineligible for copyright protection. However, they can be protected under the *sui generis* database right.[1061] Such a protection is granted to databases that involve "substantial investment in either the obtaining, verification or presentation of the contents," even if there is no originality in that database.[1062] In practice, achieving genuine creativity in a database schema proves difficult, and courts impose a high standard for "substantial investment." For instance, the CJEU ruled that Article 7 of the Database Directive should be interpreted to mean that the creator of a database can prohibit internet search engines from extracting and re-utilizing content if such actions negatively impact the creator's investment in obtaining, verifying, or presenting that content. In other words, for the database to be protected, these acts must constitute a risk to the possibility of redeeming the maker's investment through the normal operation of the database in question.[1063]

b) The necessary permission from rights holders

For content protected by intellectual property rights, obtaining consent from the rights holders is necessary for any utilization or reproduction. Training an AI system is not just a temporary reproduction; it involves the long-term ingestion and processing of data. Therefore, it is necessary for AI developers to obtain explicit permission from the rights holders in order to use their material for training. In practice, this condition is difficult, if not impossible, to meet. The vast scale of datasets involved and the multitude of rights holders concerned make it highly impractical for those training AI models to seek and obtain permission from each individual rights holder or website owner. Certainly, online content (text, images) is sometimes subject to permissive licensing terms. For example, Creative Commons licenses permit reproduction and reuse of the content, including for commercial purposes. However, the use of such licenses is not a widely available option.

Furthermore, even if website owners are not protected against unauthorized use of their data based on intellectual property rights, they can still impose contractual restrictions to prevent other businesses from

---

[1060] Database Directive, art. 3.

[1061] *Id.* art. 7.

[1062] art. 7(1) of the Database Directive mentions that the protection applies if "the maker of a database (...) shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database."

[1063] Court of Justice of the European Union (CJUE), case C-762/19, CV-Online Latvia SIA v Melons SIA (June 3, 2021) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62019CJ0762; *see* Martin Husovec & Estelle Derclaye, *Access to information and competition concerns enter the sui generis right's infringement test – The CJEU redefines the database right*, KLUWER COPYRIGHT BLOG (June 7, 2021), https://copyrightblog.kluweriplaw.com/2021/06/17/access-to-information-and-competition-concerns-enter-the-sui-generis-rights-infringement-test-the-cjeu-redefines-the-database-right/.

scraping information from their sites. The CJEU has ruled in this way in the case of *Ryanair Ltd v. PR Aviation BV*.[1064] PR Aviation operated a flight aggregation service and scraped Ryanair's data to display its flights within their search results. Ryanair filed a lawsuit to halt this practice. The court decided that Ryanair's data were not eligible for protection under copyright or a sui generis right. However, the court acknowledged that Ryanair could restrict scraping through its terms of service. Therefore, website owners can include contractual clauses within their website's Terms and Conditions that forbid web scraping, even if some or all content on their website lacks inherent protection under intellectual property rights. This is frequent in practice.

c) The Text and Data Mining exception

A potential solution to ensure the lawful use of web scraped data to train generative AI models would be to rely on the Text and Data Mining (TDM) exception provided by Articles 3 and 4 of the 2019 Copyright Directive ("New Copyright Directive").[1065] Article 2(2) of the New Copyright Directive defines Text and Data Mining as "any automated analytical technique aimed at analyzing text and data in digital form to generate information which includes but is not limited to patterns, trends and correlations."

Article 3 of the New Copyright Directive provides that TDM activities conducted by "research organisations and cultural heritage institutions" are authorized. Article 2(3) defines a "cultural heritage institution" as "a publicly accessible library or museum, an archive or a film or audio heritage institution." According to Article 2(1), a "research organisation" is either a not-for-profit entity or an entity tasked by a Member State with a public service research mission. Article 3 permits TDM only in respect of works or other subject matter (e.g., databases) to which beneficiary organizations "have lawful access." According to Recital 14, "lawful access" covers access to content pursuant to contractual arrangements (e.g., subscriptions or open access licenses), as well as to "content that is freely available online." However, Article 3 does not apply when TDM activities are carried out by private companies or for commercial motives.

Article 4(1) of the New Copyright Directive permits the "reproductions and extractions of lawfully accessible works and other subject matter for text and data mining purposes." The provision thus permits TDM for all imaginable purposes. Until now, there was a general consensus that the Text and Data Mining (TDM) exception covers the use of copyrighted works for training AI models. The explicit reference to Article 4(3) of the New Copyright Directive in Article 53(1)(c) of the AI Act (*see below 5.1.2.C.1.*) confirms this interpretation.[1066]

[1064] Court of Justice of the European Union (CJUE), case C-30/14, Ryanair v PR Aviation BV, (Jan. 15, 2015), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62014CJ0030.

[1065] Directive (EU) 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market (**New Copyright Directive**), O.J. (L 130, 17.5.2019) http://data.europa.eu/eli/dir/2019/790/oj; *see* Joao Pedro Quintais, *Generative AI, Copyright and the AI Act*, KLUWER COPYRIGHT BLOG (May 9, 2023), https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/.

[1066] Paul Keller, *A first look at the copyright relevant parts in the final AI Act compromise*, KLUWER COPYRIGHT BLOG (Dec. 11, 2023), https://copyrightblog.kluweriplaw.com/2023/12/11/a-first-look-at-the-copyright-relevant-parts-in-the-final-ai-act-compromise/.

Article 4(2) of the New Copyright Directive states that the reproductions and extractions of content made under Article 4(1) may be retained "for as long as is necessary for the purposes of text and data mining." This seems to imply that copyrighted content used during training should be deleted immediately after training. To avoid such a problematic consequence, some scholars promote "a broad normative interpretation of 'text and data mining,' encompassing not only the training activity in the strict sense but also the validation and testing" of the model.[1067]

d) Limitations to the TDM exception

The TDM exception covers only cases where content was lawfully accessed, which includes content "freely available online"[1068] and content accessed pursuant to contractual arrangements (e.g., subscriptions or open access licenses). It is also required that the use of protected materials "has not been expressly reserved by their right-holders in an appropriate manner, such as machine-readable means in the case of content made publicly available online," as outlined by Article 4(3) of the Directive.

Hence, rights holders can establish a legally recognized reservation of use in a format readable by machines. According to Recital 18 of the New Copyright Directive, "it should only be considered appropriate to reserve those rights by the use of machine-readable means, including metadata and terms and conditions of a website or a service. […] In other cases, it can be appropriate to reserve the rights by other means, such as contractual agreements or a unilateral declaration." In this context, the tools typically employed—usually web crawlers[1069]—to compile the extensive datasets required for training should be equipped to autonomously analyze websites' metadata.[1070] This functionality allows them to differentiate between materials where usage rights have not been explicitly reserved by their rights holders and materials that are freely accessible for training purposes. For example, OpenAI's GPTBot[1071] is designed to filter out unwanted sources and customize access: GPTBot identifies itself so web owners can block it via robots.txt files.[1072]

Moreover, the TDM exception should, in principle, apply only "in certain special cases which do not conflict with a normal exploitation of the work or other subject matter and do not unreasonably prejudice

---

[1067] Novelli et al., *supra* note 1039, at 16.

[1068] New Copyright Directive, Recital 14.

[1069] These tools are automated programs or bots that systematically search websites and index the content on them. They are used to scrape or pull content from websites.

[1070] Novelli et al., *supra* note 1039, at 15.

[1071] *GPTBot*, OPENAI, https://platform.openai.com/docs/gptbot (last visited Feb. 23, 2024).

[1072] Historically, website operators have used these files to signal to search engines that their sites should not be indexed. However, robots.txt files lack the ability to discriminate: If a website owner specifies that their site should not be scraped for training purposes, it will also be omitted from search engine results. As a result, the online content effectively vanishes from online visibility. *see* Katharina de la Durantaye, *Garbage In, Garbage Out. Regulating Generative AI Through Copyright Law*, ZUM (Oct. 2023) 645–60 at 10, 19; https://ssrn.com/abstract=4572952 or http://dx.doi.org/10.2139/ssrn.4572952.

the legitimate interests of the rightholder."[1073] The question, therefore, is whether using web scraped content to integrate it into datasets for training generative AI models constitutes "normal exploitation" of this content. It is also essential to determine whether such use could "unreasonably prejudice" the interests of rights holders. This situation could arise when models trained on copyrighted works produce content that diminishes public interest in the original sources.[1074] However, such harm would be difficult to establish.

e) Litigation

In the European Union, the use of copyrighted works by developers of generative AI has given rise to fewer disputes than in the United States.[1075] Recent litigation includes the case concerning the LAION database and the recent sanction of Google.

i) The LAION case

In Germany, a photographer filed a lawsuit against LAION e.V. (Large-scale Artificial Intelligence Open Network), a nonprofit organization that curates training datasets for developers of generative AI models.[1076] LAION trained two AI models, LAION-400M and LAION-5B, with publicly available images. These datasets are widely used for training the most popular generative AI tools: among others, Stability AI utilized LAION 5B for training the Stable Diffusion model.

The photographer alleged that his photos were included in the LAION 5B dataset and requested LAION to remove them. Subsequently, he filed for an injunction to prevent the use of his images by LAION, asserting that such use constitutes copyright infringement through data mining. LAION responded that its databases do not store pixel data, but only plain text data, metadata, and URLs, which LAION-400M and LAION-5B use to link to images available elsewhere on the Internet. According to LAION, the datasets serve as index directories for locating image material on the free Internet. While links to specific images can be removed from the catalog, the actual images cannot be removed, as the association does not store images in its databases.

This litigation is currently pending. The discussion will focus on the potential interpretation of German Copyright Law.[1077] The German Act on Copyright and Related Rights (*Urheberrechtsgesetz*) includes two exceptions for text and data mining that LAION invoked when creating its datasets. Section 44b of the

---

[1073] Copyright Directive, art. 5(5).

[1074] According to Novelli et al., "the TDM exception cannot justify reproductions that lead to applications that substitute, or otherwise significantly economically compete with, the protected material used for AI training. However, this is, arguably, precisely what many generative AI applications are doing." Novelli et al., *supra* note 1039, at 16.

[1075] Durantaye, *supra* note 1072, at 1.

[1076] *An Up-Date on the Robert Kneswhke v. LAION e.V lawsuit,* CEPIC (Nov. 23, 2023) https://cepic.org/news/an-up-date-on-the-robert-kneschke-v-laion-e-v.

[1077] Silke Hahn, *Stock photographer sues AI association LAION: The crux with AI training data*, HEISE ONLINE (May 5, 2023), https://www.heise.de/hintergrund/Stock-photographer-sues-AI-association-LAION-The-crux-with-AI-training-data-8988690.html.

German Copyright Act provides a general text and data mining exception, as long as the image data is used solely for pattern recognition or analysis and the image-text pairs are not stored after evaluation. Section 60d, which is more specifically tailored, provides an exception for research purposes, provided the results are not intended for commercial use or that any revenue generated is reinvested into research. While this litigation should clarify the rights of image creators whose works are used to train AI models, it is important to note that LAION's nonprofit status is a crucial aspect of this case.

ii) The sanction of Google by the French Competition Authority

The French Competition Authority fined Google €250 million on March 20, 2024, for noncompliance with the New Copyright Directive.[1078] Article 15 of this Directive introduces a new related right for EU-based press publishers by granting them the right to earn revenue from the online use of their publications by information society service providers.[1079] In addition, providers must pay the authors of press publications an appropriate share of the revenues obtained from the licensing of online uses of their works. Additionally, Article 16 of the New Copyright Directive grants all publishers, including those of press publications, books, music, and scientific works, the right to receive fair compensation. The French law of 24 July 2019 implemented these provisions.[1080]

The French Competition Authority has chosen to sanction Google on several grounds, many of which pertain to Bard (now named Gemini) using content from press agencies and publishers to train its foundation model. Firstly, the French Competition Authority concluded that Google's failure to notify editors and press agencies about the use of their content violated the transparency obligations to which Google had previously committed. This lack of communication impaired the ability of press agencies and publishers to negotiate for fair remuneration. Secondly, the Authority found that Google did not uphold its commitment to ensure the neutrality of negotiations regarding related rights, separate from any other economic relationships it may have with publishers and press agencies.

Indeed, up until at least September 28, 2023, Google failed to provide a technical solution that would enable publishers and press agencies to opt out of having their content used to train Bard without affecting the display of other content protected by related rights on Google's services. Publishers and news agencies had the option to add a "noindex" tag to the robots.txt file. This file, located in the root directory of web servers, contains directives for search engine crawlers. However, using a "noindex" tag resulted in the complete removal of a website from Google's search results, including Search, Discovery, and Google News, which

[1078] Autorité de la Concurrence, Décision 24-D-03 (March 15, 2024) https://www.autoritedelaconcurrence.fr/fr/decision/relative-au-respect-des-engagements-figurant-dans-la-decision-de-lautorite-de-la-0; Natasha Lomas &, Romain Dillet, *Google hit with $270M fine in France as authority finds news publishers' data was used for Gemini*, TECHCRUNCH, (Mar. 20, 2024) https://techcrunch.com/2024/03/20/google-hit-with-270m-fine-in-france-as-authority-finds-news-publishers-data-was-used-for-gemini/?guccounter=1.

[1079] Information Society Providers are defined in EU Law as "any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services", pursuant to art. (1)(1)(b) of Directive 2015/1535/EU.

[1080] Loi n° 2019-775 du 24 juillet 2019 tendant à créer un droit voisin au profit des agences de presse et des éditeurs de presse (Law No. 2019-775 of July 24, 2019, aimed at creating a related right for the benefit of news agencies and press publishers)

were precisely the subject of negotiations for the remuneration of related rights. By doing so, Google tied the use of publishers' and news agencies' content for training its AI models to the display of protected content. That eliminated the ability of publishers and news agencies to negotiate remuneration. In September 2023, Google introduced more nuanced options, including a new "Google-Extended" rule that allows web publishers to specify their preference to exclude their content from being used to train AI models.

2) Copyright infringements by AI-generated outputs

Generative AI models are sometimes referred to as "stochastic parrots,"[1081] as they are suspected to duplicate or simply regurgitate content in their training datasets. In principle, it is relatively uncommon for well-trained, state-of-the-art generative AI models to outright duplicate existing content. Generative AI models primarily focus on learning overarching patterns and styles to generate *new* content. Sometimes, however, an AI model may simply regurgitate content present in its training data. This is what Github Copilot has been accused of doing in pending litigation in the U.S (*see above section 3.3.2.B)*.[1082]

Determining whether output from generative AI constitutes copyright infringement cannot be conclusively addressed in the abstract.[1083] To reach a definitive conclusion, it is necessary to do a thorough examination on a case-by-case basis, comparing the AI-generated output to the prior copyrighted works. If the AI-generated output shows substantial and direct similarities to legally protected elements of existing materials, it is probable that the AI-generated work would be seen as infringing upon the copyrights of those prior copyrighted materials. Furthermore, should the output incorporate protected aspects or elements of existing materials, it would likely be considered as a derivative creation based on those pre-existing materials.[1084]

Once copyright infringement is established, the question shifts to determine liability for the infringement: Does liability fall to the individual user who input the request or does it fall on the provider of the AI tool that generated the output? This inquiry mirrors the one that arises in cases involving the creation of illicit content, such as hate speech, as defined by the legislation of Member States. This question is addressed in section 5.1.1.C.

3) Copyrightability or patentability of generated outputs

Generative AI models have the capacity to generate ideas or designs that, if conceived and produced by a human alone, could qualify for intellectual property protection. However, because generative AI typically relies on human inputs to generate the AI model's outputs, the AI model may be perceived as a *tool*, not an autonomous *creator* or inventor. The determination of whether outputs from generative AI *can* be protected by intellectual property (IP) law requires an analysis of its potential eligibility for copyright (in the case of literary, dramatic, musical, or artistic works) and patent protection (in the case of inventions).

[1081] Bender et al., *supra* note 221.

[1082] Compl., J.Doe 1 v. Github, Inc., No. 3:22-cv-06823 (N.D. Cal. Nov. 3, 2022).

[1083] Novelli et al., *supra* note 1039, at 17.

[1084] *Id.*

a) Copyrightability

As for copyright protection, EU legislation does not explicitly state that the "author" must be human. For a work to be eligible for protection, it must be original —that is, it must constitute the author's intellectual creation. The CJEU has not specifically addressed the copyright status of AI-generated works, but it has established that copyright protection requires input reflecting the author's personality.[1085] This seems to imply that such input must be from a human.

Certainly, it is possible to produce copyright-protected works with the assistance of an AI device. However, even if works can be AI-assisted, they must meet the criteria of originality and creativity. This is why there may be hesitation when human intervention is limited to providing a prompt or when only minor alterations are made to the output generated by the AI model, such as minor edits to a machine-generated text.[1086] In any case, works entirely generated by a computer, without any human involvement, do not appear to be eligible for copyright protection.

b) Patentability

To gain a patent in Europe, an invention must be new, involve an inventive step, and be applied to an industrial application.[1087] As with copyrights, the patent protection of inventions created by a person with the assistance of AI is possible.[1088] However, the European Patent Office (EPO) has ruled that, under the European Patent Convention, an inventor designated in a patent application must be "a person with legal capacity."[1089] In other words, an AI system cannot be recognized as an inventor on a patent application. Only human beings can be listed as inventors. Inventions that are autonomously created by an AI system, without any human involvement, will not be attributed to any inventor, including the owner of the AI model and, hence, the invention cannot be patented.

[1085] Court of Justice of the European Union (CJUE), case C-833/18 SI, Brompton Bicycle Ltd v Chedech/Get2Get (June 11, 2020), https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:62018CJ0833; *see* Johannes Fritz, *The notion of 'authorship' under EU law —who can be an author and what makes one an author? An analysis of the legislative framework and case law*, JOURNAL OF INTELLECTUAL PROPERTY LAW & PRACTICE, 2024, https://doi.org/10.1093/jiplp/jpae022.

[1086] Novelli et al., *supra* note 1039. P. Bernt Hugenholtz & João Pedro Quintais, *Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?*, INTERNATIONAL REVIEW OF INTELLECTUAL PROPERTY AND COMPETITION LAW 52, 1190–1216 (2021), https://link-springer-com.stanford.idm.oclc.org/article/10.1007/s40319-021-01115-0#citeas.

[1087] European Patent Convention, art. 5, https://www.epo.org/en/legal/epc/2020/a52.html.

[1088] This probably implies the adoption of a broad interpretation of the inventive step requirement, *see* Novelli et al., *supra* 1039.

[1089] On December 21, 2021, the Legal Board of Appeal of the EPO issued a decision in case J8/20, ruling that under the European Patent Convention, an inventor designated in a patent application must be "a person with legal capacity," European Patent Office, case J0008/20-3.1.01, Stephen L. Thaler, (Jan. 3, 2022) https://register.epo.org/application?documentId=KXGBKNEA11IZE8D&number=EP18275163&lng=en&npl=false .

5.1.1.C. Liability for machine-generated content

There are various types of liability concerns possible with generative AI. One specific concern emerges when an AI model's performance is suboptimal or faulty. For example, a model may display a technical vulnerability that leads to the replication of training data, as demonstrated by the example of ChatGPT regurgitating three and a half chapters from *Harry Potter and the Sorcerer's Stone*.[1090] An AI tool may also malfunction and be incapable of performing its typical functions. ChatGPT has recently exhibited atypical behavior, switching between languages and making nonsensical responses.[1091] In the latter two scenarios, the AI model did not perform as designed. In such cases, AI companies should be ready to demonstrate that their AI model has been rigorously designed and tested. If they fail to do so, they could be considered at fault and, therefore, responsible for the malfunction. The prospect of being held liable for any dysfunction or malfunction should motivate AI developers to ensure their software is crafted to the highest standards of quality and reliability.

Another concern arises when a generative AI tool functions as intended but produces outcomes that are potentially unlawful (such as an erroneous statement about an individual that turns out to be defamatory) or clearly unlawful (such as incitement to commit a violent act). In this regard, the extensive adoption of generative AI models will almost certainly raise new questions about the legal definition of unacceptable speech. For example, legal provisions may stipulate that the speaker's *intention* must be taken into consideration when assessing whether the person's speech is unlawful. Yet an AI tool lacks intentionality. Moreover, AI-generated content that is not illegal could still cause harm if disseminated, as is the case with misinformation or disinformation.

The following sections will concentrate on examining liability issues arising from the generation of content that violates legal norms. In the European Union, the principles of civil liability and the legal provisions defining illegal content are generally governed by the domestic law of each Member State. Most European countries adhere to a general principle of fault-based liability. And most Member States have comprehensive legal frameworks to regulate prohibited speech and content, encompassing categories such as defamation, hate speech, child abuse material, incitement to violence, and, sometimes, disinformation.

So, who *is* liable when an AI model generates unlawful content? The end user or the provider of the generative AI system?

1) Liability of end users

A useful place to begin answering this question is with a scenario that is easy to imagine happening: An individual user obtains illicit content after prompting a generative AI chatbot to create it. There are several arguments to support the conclusion that liability lies with this individual user.

[1090] Peter Henderson et al., *Foundation Models and Copyright Questions*, STANFORD HAI (Nov. 2023) at 8, https://hai.stanford.edu/sites/default/files/2023-11/Foundation-Models-Copyright.pdf.

[1091] Wes Davis, *ChatGPT spat out gibberish for many users overnight before OpenAI fixed it*, THE VERGE, (February 21, 2024), https://www.theverge.com/2024/2/21/24079047/chatgpt-malfunction-hallucination-responses-openai.

First, it could be argued that, when the generative AI system generates illegal output, the user should abstain from sharing or reproducing it. Providers of generative AI services generally warn users of the potential for the models to produce inaccurate or misleading outputs. The chatbot interfaces of generative AI companies typically incorporate warnings and disclaimers.[1092] However, it is not always apparent to an average user that a specific piece of content is unlawful. For instance, consider a text generated by a generative AI chatbot that falsely implicates an individual in criminal activities. The user may not necessarily recognize these statements as false and, thus, defamatory.

Second, the creation of illegal content with a generative AI chatbot primarily results from the actions of the user who prompts the model. However, it is difficult to determine the extent to which the *output* is influenced by the *user's input*, rather than the independent operation of the model, churning through data and assembling a response. Certainly, if the user's input appears explicitly *aimed* at generating an unlawful outcome —such as hate speech or child sexual abuse material— it is plausible to argue that the user is committing a civil wrong. Similarly, if the user inputs a prompt that violates the AI providers' terms of service, the user is in contractual breach and liable if the resulting output is illegal. However, in all the hypotheticals mentioned, assessing the user's liability requires access to the content of their prompts, which may be difficult to obtain. Moreover, finding the user liable does not necessarily preclude the possibility that the provider will also be liable.

2) Liability of providers

The first thought regarding the liability of generative AI service providers is whether they benefit from the liability exemption reserved for hosting providers in EU law. However, this is not the case. Instead, they are primarily subject to the general liability regime outlined by each national law.

a) No liability exemption for generative AI providers

In 2000, the EU Directive 2000/31/EC (The E-Commerce Directive)[1093] introduced a new category of service providers in EU law: "hosting service providers." According to Article 14 of the Directive, "hosting service providers" are exempt from liability for content stored at a user's initiative, as long as the providers had no actual knowledge of the content's illegality. Similarly, providers of mere conduit and caching services are not held liable for the information they transmit or store for their users under the same conditions.

The recently adopted Digital Services Act (DSA) maintains this knowledge-based liability principle.[1094] It continues to grant immunity to these service providers, provided they act "expeditiously" to remove or

[1092] *Terms of Use*, OPENAI, https://openai.com/fr-FR/policies/terms-of-use/ (last visited June 20, 2024)

[1093] Directive 2000/31/EC of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (**Directive on electronic commerce**) O.J. (L 178, 17.7.2000), http://data.europa.eu/eli/dir/2000/31/oj.

[1094] Council Regulation (EU) 2022/2065 of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (**Digital Services Act**) O.J. (L 277, 27.10.2022), http://data.europa.eu/eli/reg/2022/2065/oj; *See* Florence G'sell, *in* Antje von Ungern-Sternberg (ed.), *Content Regulation in the European Union – The Digital*

disable access to illicit content once they become aware of its illegality. Specifically, Article 6 of the DSA grants the liability exemption to any information society service provider whose service consists of "the storage of information provided by a recipient of the service."[1095] The liability exemption also benefits any provider whose service "consists of the transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network,"'[1096] or "of the transmission in a communication network of information provided by a recipient of the service."[1097]

Given the clear and precise terms of the DSA provisions, it is evident that providers of generative AI services cannot benefit from the liability exemption, as they do not offer the types of services targeted by these provisions. This exclusion is logical, since providers of generative AI services do not aim to host or transmit content. Generative AI tools produce content based on user instructions but are not designed to store and disseminate content created by third parties.

b) The traditional fault-based liability regime

For now, providers of generative AI services are primarily subject to the general liability regime established by the national laws of European Member States, which is typically fault-based (to be distinguished from product liability *(see below section 5.1.3.A))*.

In this traditional framework of fault-based liability, it does not seem that the provider of an AI system can be held civilly liable solely because the system has generated illegal content. In order to prove civil fault, it is typically required to demonstrate that the defendant did not comply with a general standard of reasonable, cautious, and diligent conduct. Confronted with exceptionally sophisticated and frequently unpredictable generative AI systems, it will be difficult to establish that the provider has *not* adhered to such a standard. This may require proving that the model was trained using illicit content or that the developer or deployer implemented no guardrails to avoid the generation of illegal material. In any case, it would likely be highly challenging for a plaintiff to obtain this evidence, especially since most European procedural systems do not have any process similar to the civil procedure of discovery used in the US.

The question of liability could become even more difficult within complex supply chains *(see above section 2.3)*. It is routine for a single AI model to be developed, fine-tuned, and used by multiple parties along the supply chain. Third parties may leverage APIs to access foundation models created by others to build new downstream applications onto them. Certainly, the situation is more straightforward when an AI model is provided by a specific provider who establishes a contractual agreement with the model's deployers and users. However, complexity and uncertainty increase with open-source foundation models, which are subject to modifications by numerous contributors.

Services Act, SCHRIFTEN DES IRDT - TRIER STUDIES ON DIGITAL LAW, Volume 1, *Verein für Recht und Digitalisierung e.V.*, INSTITUTE FOR DIGITAL LAW TRIER (April 2023), https://ssrn.com/abstract=4403433 or http://dx.doi.org/10.2139/ssrn.4403433.

[1095] Digital Services Act, art. 6 (Hosting).

[1096] *Id*. art. 4 (Mere conduit),.

[1097] *Id.* art. 5 (Caching).

5.1.1.D The obligations of the Digital Services Act

Numerous European regulations may impact providers of generative AI systems by imposing various obligations on them. Among the various applicable regulatory frameworks, the recently enacted Digital Services Act Regulation (EU) 2022/2065 (DSA)[1098] warrants particular attention due to its focus on mitigating the spread of harmful speech. Although the DSA is not explicitly crafted to regulate developers and providers of generative AI, its provisions are relevant to generative AI models and systems that are integrated into the moderation tools employed by platforms and search engines.

1) Generative AI providers fall outside the scope of the DSA

The Digital Services Act does not merely uphold the liability exemption previously provided by the E-commerce Directive. It also imposes a number of obligations on "intermediary services." However, as stated in the previous paragraph *(see section 5.1.1.C(2)*), generative AI services do not directly fit into the categories of services regulated by the DSA.[1099] "Intermediary services" are defined in Article 3(g) as "mere conduit[s]" (such as internet access providers), "caching" services (storage for quick retrieval of files), or "hosting" services (such as social media platforms). These services only transmit, store, or host content provided by users,[1100] which differs fundamentally from the generation of content based on user requests.[1101] While the liability exemption applies only to the three aforementioned categories of "intermediary services," the DSA also imposes obligations on two new categories: "online platforms,"[1102] which are a category of hosting services that disseminate information to the public, and "online search engines."

The question then arises whether generative AI services could be classified as "search engines" and thus be subject to the corresponding obligations, which primarily apply to "Very Large Search Engines" with at least 45 million users. The DSA defines "online search engine" as an "intermediary service that allows users to input queries in order to perform searches of, in principle, all websites, or all websites in a particular language, on the basis of a query."[1103] "The query can be on any subject and in the form of a keyword, voice request, phrase or other input," states the DSA, "and the AI model returns results in any format in which information related to the requested content can be found."[1104] Generative AI systems are, indeed, designed to address user queries. And their operation entails diverse interactions between third-party content, such as training datasets and internet search results, and the outputs of the system. However, the primary function

---

[1098] *Id.*

[1099] Philipp Hacker, et al., *Understanding and Regulating ChatGPT, and Other Large Generative AI Models: With Input from ChatGPT*, VERFASSUNGSBLOG (Jan. 20, 2023), https://verfassungsblog.de/chatgpt/

[1100] For example, "hosting providers" store information provided by, and at the request of, users, as outlined by art. 3(g)(iii) of the DSA.

[1101] Philipp Hacker et al., *Regulating ChatGPT and other Large Generative AI Models*, FACCT'23, June 12-15, 2023, Chicago, IL, USA, at 1118.

[1102] art. 3(i) of the DSA provides that hosting providers are hosting services that store and disseminate information to the public, as social networks or marketplaces.

[1103] DSA, art. 3(j).

[1104] *Id.*

of generative AI tools is not to conduct web searches to retrieve and organize existing web-based information and guide users to the sources that best match their search criteria.

Of course, generative AI systems could be seen as operating on "a spectrum between a retrieval search engine (...) and a creative engine.[1105] In the US, some scholars have argued that, even when their outputs seem novel or creative, they are still ultimately dependent on third-party content from training data and user prompts.[1106] Nonetheless, the ability of generative AI models to generate original, substantial content from brief user prompts undermines the argument that their outputs are simply information sourced from another information content provider. Although it is accurate to say that these models undergo training using data scraped from the web, their core purpose is to *generate new and original* content based on learned patterns and knowledge, rather than merely performing search operations.

2) DSA provisions applicable to VLOPs and VLOSEs

While the Digital Services Act does not directly regulate generative AI systems, it indirectly applies to them. This occurs because the DSA governs "Very Large Online Platforms" (VLOPs) and "Very Large Online Search Engines" (VLOSEs) with at least 45 million users. These VLOPs and VLOSEs use generative AI models as part of their content curation or moderation tools.[1107]

VLOPs and VLOSEs must report annually on their content moderation, as outlined by Article 15 of the DSA. They must publish transparency reports twice a year that include information about their content moderation resources, as provided by Article 42. In addition, they must provide regulators with access to the data needed to verify that they are in compliance with the DSA, as stated by Article 40(1). In particular, they must be able to explain to regulators the design, logic, operation, and testing of their algorithmic systems, including their recommendation systems. Of particular note is the fact that Article 15(1)(e) requires "a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error" of the automated tools used. VLOPs and VLOSEs must be able to explain to regulators "the design, logic, operation, and testing of their algorithmic systems, including their recommendation systems."[1108]

Furthermore, Article 34 of the DSA provides that VLOPs and VLOSEs must evaluate and address "systemic risks" through appropriate policies. They must analyze the extent to which their moderation, recommendation, and advertising systems may affect those systemic risks. This should be done annually and also prior to the deployment of functionalities that are likely to have a critical impact on the risks identified. Systemic risks pertain to issues such as illegal content, hate speech, privacy violations, election

---

[1105] P. Henderson et al., *Where's the Liability in Harmful AI Speech?*, Social Science Research (August 9, 2023), p. 622, https://www.journaloffreespeechlaw.org/hendersonhashimotolemley.pdf.

[1106] Derek E. Bambauer & Mihai Surdeanu, *Authorbots* (May 9, 2023) 3 Journal of Free Speech Law (Forthcoming), Arizona Legal Studies Discussion Paper No. 23-13, https://ssrn.com/abstract=4443714; Beatriz Botero Arcila, *Is it a Platform? Is it a Search Engine? It's ChatGPT! The European Liability Regime for Large Language Models* (Aug. 12, 2023), Journal of Free Speech Law, Vol. 3, Issue 2, 2023, https://ssrn.com/abstract=4539452.

[1107] Lilian Weng et al., *Using GPT-4 for Content Moderation*, OpenAI, (Aug. 15, 2023) https://openai.com/index/using-gpt-4-for-content-moderation/.

[1108] DSA, art. 40(3).

manipulation, and other similar problems. Moreover, content that generates adverse effects on fundamental rights, civic discourse, electoral processes, public security, gender-based violence, public health, minors, and personal well-being may also lead to systemic risks. Although Article 35(1) mentions "illegal hate speech or cyber violence," the definition of systemic risks encompasses content that is not necessarily illegal but may cause problems, such as misinformation on public health, climate change, or politics.

After assessing systemic risks, VLOPs and VLOSEs must implement "reasonable, proportionate, and effective mitigation measures" to counter such risks, as provided by Article 35 of the DSA. These measures may include adapting the design of the interfaces, adapting the terms and conditions, improving the notice and action mechanism, improving the algorithmic systems, increasing the visibility of reliable information sources, labeling suspicious content, or implementing codes of conduct.[1109] Upon request, VLOPs and VLOSEs must provide to the European Commission and relevant national Digital Services Coordinators their assessments of systemic risks.[1110] And, the Commission, in cooperation with Digital Services Coordinators, may issue guidelines and recommend actions.[1111]

Within this framework, the EU Commission included "generative models" in its Delegated Regulation,[1112] laying down rules on the performance of audits for VLOPs and VLOSEs under the DSA. Recently, the EU Commission issued requests for information to six VLOPs and two VLOSEs, including Bing, Facebook, Google Search, Instagram, Snapchat, TikTok, YouTube, and X.[1113] These requests aim to gather detailed information on how these platforms are addressing risks associated with generative AI in their services. Specifically, the Commission seeks to understand measures taken to mitigate issues such as "hallucinations" (instances where AI generates false information), the widespread distribution of deepfakes, and the automated manipulation of services that could deceive voters.

Additionally, the Commission inquires about risk assessments and mitigation strategies related to generative AI's impact on electoral processes, the spread of illegal content, the protection of fundamental rights, gender-based violence, the safety of minors, and mental health. The inquiries encompass both the dissemination and the creation of content by generative AI systems. The EU Commission published guidelines that include examples of potential mitigation strategies for election-related risks.[1114] These

---

[1109] DSA, art. 35(1).

[1110] *Id.* art. 35(2).

[1111] *Id.* art. 35(3).

[1112] Commission Delegated Regulation (EU) 2024/436 of 20 October 2023 supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council, by laying down rules on the performance of audits for very large online platforms and very large online search engines, O.J. (L, 2024/436, 2.2.2024), http://data.europa.eu/eli/reg_del/2024/436/oj.

[1113] *Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act***,** EUROPEAN COMMISSION, (Mar. 14, 2024), https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very#:~:text=The%20European%20Commission%20formally%20sent,%3A%20such%.

[1114] *Commission publishes guidelines under the DSA for the mitigation of systemic risks online for elections*, EUROPEAN COMMISSION, (Mar. 26, 2024) https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1707.

guidelines encompass measures to mitigate the risks posed by generative AI content, for example, by clearly labeling AI-generated content (such as deepfakes) or adapting the platforms' terms and conditions.

5.1.2 The AI Act

The concept of adopting comprehensive legislation specifically targeting AI has gradually gained traction in the EU. The European Commission's 2018 Communication "Artificial Intelligence for Europe"[1115] underscored the importance of establishing a suitable ethical and legal framework aligned with the European Union's values and the Charter of Fundamental Rights. However, it did not advocate for the creation of new, binding legislation specifically tailored to artificial intelligence. Instead, the Communication directed attention to the existing regulatory frameworks, particularly those concerning personal data protection, product safety, and civil liability. It emphasized that these frameworks could provide a foundation for further development.

Following this Communication, the European Commission established an independent High-Level Expert Group on Artificial Intelligence (AI HLEG)[1116] to define criteria for "Trustworthy AI." After thorough discussions, the group unveiled a set of seven key requirements for Trustworthy AI:

- human agency and oversight;
- technical robustness and safety;
- privacy and data governance;
- transparency;
- diversity, nondiscrimination, and fairness;
- societal and environmental well-being; and
- accountability.[1117]

The Commission validated these principles in its April 2019 document, "Building Trust in Human-Centric Artificial Intelligence."[1118] In July 2020, the expert group introduced a practical tool for applying these principles: the Assessment List for Trustworthy AI (ALTAI).[1119]

[1115] European Commission, *Communication: Artificial Intelligence for Europe*, {SWD(2018) 137 final} COM/2018/237 final, (April 25 2018), EUROPEAN COMMISSION, https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe.

[1116] *High-level expert group on artificial intelligence*, EUROPEAN COMMISSION, https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (last visited June 20, 2024).

[1117] AI HLEG, *Ethics guidelines for trustworthy AI*, EUROPEAN COMMISSION (Apr. 8, 2019) https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[1118] European Commission, *Communication: Building Trust in Human Centric Artificial Intelligence* (COM(2019)168), EUROPEAN COMMISSION, (Apr. 8, 2019), https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence.

[1119] AI HLEG, *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*, EUROPEAN COMMISSION (June 17, 2020), https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

The impetus for proposing the AI Act originated from the political guidelines issued by (then candidate for) EU Commission President Ursula von der Leyen in July 2019.[1120] The guidelines demanded a unified European approach to the human and ethical challenges posed by artificial intelligence. A Commission white paper,[1121] released in February 2020, offered policy guidance for crafting a regulatory framework and an investment strategy. The white paper outlined policy options on how to achieve the dual objectives of promoting the uptake of AI and addressing the risks associated with certain uses of the technology. It advocated for a risk-based approach and insisted on two primary issues: the prevention of violations of fundamental rights through AI use, and the enhancement of the liability framework's effectiveness.

The European Union's proposed Regulation on Artificial Intelligence[1122] (also known as the AI Act) was unveiled by the European Commission on April 21, 2021. This proposal formed part of a broader, strategic initiative to shape the EU's digital economy over the next decade.[1123] While the drafting of the AI Act was taking its course from the Commission's proposal, the EU Commission published, on September 28, 2022, a Proposal for an AI Liability Directive[1124] and a Proposal for a revision of the Product Liability Directive.[1125] In the following months, the European Parliament, the EU Council, and the European Commission adopted the "European Declaration on Digital Rights and Principles for the Digital Decade."[1126] This European Declaration, adopted in December 2022, is a non-binding document but one that articulates a set of principles designed to steer the EU's legislative approach in the digital sector.

The European Commission, the Council, and the European Parliament reached a political consensus on the provisions of the AI Act on December 6, 2023.[1127] The Act received approval of the Committee of

---

[1120] Ursula Von der Leyen, *A Union that strives for more - My agenda for Europe, Political Guidelines For The Next European Commission 2019-2024*, *Third guideline "A Europe fit for the digital age"*, at 13, https://commission.europa.eu/document/download/063d44e9-04ed-4033-acf9-639ecb187e87_en?filename=political-guidelines-next-commission_en.pdf.

[1121] European Commission, *White Paper on Artificial Intelligence: a European approach to excellence and trust* (Feb. 19, 2020), EUROPEAN COMMISSION, https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

[1122] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final, (April 21, 2021): https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[1123] Decision (EU) 2022/2481 of 14 December 2022 establishing the Digital Decade Policy Programme 2030, O.J. (L 323, 19.12.2022), https://eur-lex.europa.eu/eli/dec/2022/2481/oj.

[1124] Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence (**AI Liability Directive**), COM/2022/496 final (Sept. 28, 2022), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0496.

[1125] Proposal for a directive on liability for defective products, COM/2022/495 final, (Sept. 28 2022), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0495..

[1126] Declaration (EU) 2023/C 23/01 of the European Commission, European Parliament and of the Council of 23 January 2023 on Digital Rights and Principles for the Digital Decade O.J. (C 23, 23.1.2023), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:JOC_2023_023_R_0001.

[1127] For a chronological presentation of the negotiations about the AI Act, see Kai Zenner, *Digitizing Europe*, https://www.kaizenner.eu/post/aiact-part3 (last visited June 20, 2024).

Permanent Representatives to the European Union (COREPER) on February 2, 2024. It was then approved, on March 13, 2024, by the European Parliament. The final version of the AI Act was published in the Official Journal of the European Union on July 12, 2024.[1128]

This section will begin with a general overview of the AI Act, followed by an analysis of the provisions applicable to specific-purpose AI systems, categorized by risk level as outlined in the Commission's initial proposal. Next, it will examine the provisions applicable to general-purpose AI models, which emerged as a significant point of debate during the negotiations. The subsequent paragraphs will detail additional provisions of the AI Act and describe the implementation process of the Regulation.

5.1.2.A General overview

The AI Act constitutes a regulatory framework for the sale and use of artificial intelligence within the European Union. Its primary purpose is to ensure the smooth functioning of the EU single market by harmonizing AI system standards across Member States. Significantly, it marks the first comprehensive law aimed at addressing the risks of artificial intelligence. The AI Act establishes a series of requirements designed to "promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter of Fundamental Rights, including democracy, the rule of law and environmental protection," as outlined by Article 1(1) of the Act. The Regulation imposes varying obligations on AI systems based on the potential risk they pose to health, safety, and fundamental rights.

The European Commission's initial proposal for the AI Act of April 21, 2021, did not explicitly address the implications of generative AI technologies, likely due to the limited recognition of foundation and generative models at that time. Moreover, the EU Commission's proposal did not focus on the technology but centered on addressing the risks associated with AI based on its use cases, especially in specific or sensitive sectors.[1129] However, the launch of ChatGPT in November 2022 and the success of generative AI led EU lawmakers to scrutinize how these technologies could be encompassed within the regulatory scope of the EU AI Act. During the negotiations, provisions were added to regulate general-purpose (i.e., foundation) models, shifting the focus from use cases to the technology itself. As a result, the AI Act now recognizes that certain AI models inherently carry specific risks, independent of their usage in specific sectors.

1) Scope of the AI Act

The AI Act has an extensive scope, regulating providers who market AI systems or models and providers who put them into service in the European market.[1130] It does not matter whether these providers are

[1128] Council Regulation 2024/1689 of June 13, 2024, (**Artificial Intelligence Act**), 2024 O.J. (L 12.7.2024), http://data.europa.eu/eli/reg/2024/1689/oj.

[1129] Claudio Novelli,et al., *AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act* (May 31, 2023). DIGITAL SOCIETY 3, 13 (2024), https://ssrn-com.stanford.idm.oclc.org/abstract=4464783.

[1130] AI Act, art. 2(1)(a).

established or located within the European Union or in a non-EU country. If their systems or models are marketed or put to use in the EU, they fall under the Act's jurisdiction.

The Act extends its scope to include other actors in the supply chain, notably the "deployers" of AI systems. Deployers are defined as individuals or organizations that use an AI system under their authority *(see figure 15 below)*. They fall under the Act's scope provided they are established or located within the European Union.[1131] Importantly, the Act applies to any provider and deployer of an AI system established in or outside the EU if the output produced by the AI system is used within the European Union.[1132] The Act also applies to importers, distributors, and product manufacturers integrating an AI system within their product and putting their name or trademark on it, as well as authorized representatives of providers established in third countries and any affected persons located in the EU.[1133]

Within this framework, the AI Act could potentially apply to AI systems developed and employed outside the European Union, if the outputs from these systems are available and used within the EU. Consequently, the Act will have a significant extraterritorial impact, affecting numerous providers and users located outside the EU's borders. The Act provides that, prior to making their AI systems available in the EU, providers of high-risk AI systems[1134] and providers of general-purpose AI models[1135] established outside the Union shall appoint an authorized representative established in the EU.

FIGURE 15. Definitions of provider and deployer under the AI Act

	Definitions
Article 3 (3)	'**Provider**' means a natural or legal person, public authority, agency, or other body that **develops** an AI system or a general-purpose AI model—or that has an AI system or a general-purpose AI model developed—and **places it on the market** or **puts the AI system into service** under the its own name or trademark, whether for payment or free of charge.
Article 3 (4)	'**Deployer**' means any natural or legal person, public authority, agency or other body **using an AI system under its authority except where the AI system is used in the course of a personal nonprofessional activity**.

The AI Act's extensive scope of application does have certain exceptions. The Act does *not* apply to:

- individuals using AI systems **for a purely personal, nonprofessional activity**;[1136]
- AI systems or models, including their output, specifically developed and put into service for the sole purpose of **scientific research and development**;[1137]

[1131] *Id.* art. 2(1)(b).
[1132] *Id.* art. 2(1)(c).
[1133] *Id.* art. 2(1) (d, e, f, g).
[1134] *Id.* art. 22.
[1135] *Id.* art. 54.
[1136] AI Act, art. 2(10).
[1137] *Id.* art. 2(6).

- activities related **to research, testing, and development that occur before an AI system or model is placed on the market or put into service**;[1138]
- AI systems used solely for **military, defense, or national security purposes**;[1139] and
- AI systems **released under free and open-source licenses**.[1140]

However, the exemption for free and open-source systems needs to be carefully qualified. As provided by Article 2(12), free and open-source AI systems are covered by the AI Act when they are:

- integrated into prohibited AI practices;
- marketed or put into service as high-risk AI systems;
- subject to transparency obligations; or
- classified as general-purpose AI systems.

2) The AI Act's definition of AI

The European Commission's original proposal defined an AI system as "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with."[1141] The detailed Annex I included a broad spectrum of algorithmic techniques, encompassing machine-learning approaches, logic and knowledge-based approaches, (knowledge bases, expert systems, etc.), and "statistical approaches, bayesian estimation, search and optimization methods." In formulating this definition, the Commission aimed for it to be adaptable and comprehensive, ensuring it covered all potentially problematic applications.

The broad nature of this definition attracted substantial criticism for its vagueness. The inclusion of statistical methods potentially extended the Act's scope to nearly all data-analyzing computer software, even basic tools like Excel spreadsheets. Therefore, throughout the legislative process, the definition of AI underwent significant alterations and was progressively refined.

In the final version of the text, Article 3(1) defines an artificial intelligence system (AI system) as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." The final definition is in line with the OECD definition.[1142] It focuses on the generation of objective-based outputs and the capability of the AI system to interact with its environment with autonomy. However, this definition remains relatively vague.

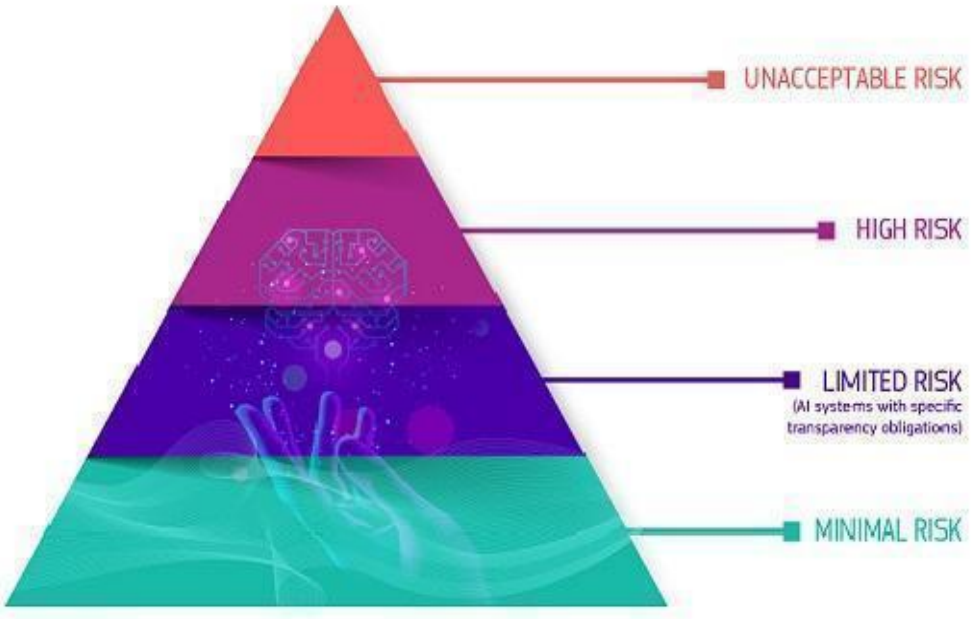
[1138] *Id.* art. 2(8).
[1139] *Id.* art. 2(3).
[1140] *Id.* art. 2(12).
[1141] *Id.* art. 3(1).
[1142] The OECD updated definition describes an AI system as "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." It specifies that "different AI systems vary in their levels of autonomy and adaptiveness after deployment." (*See above section 2.1.1).*

5.1.2.B. Specific-purpose AI systems

The drafters of the AI Act consciously chose a risk-based approach, aligning with the recommendations in the 2020 EU Commission's white paper.[1143] The aim is to align the Act's provisions and requirements with the level and range of risks that AI systems can produce. To this end, the AI Act categorizes risks based on the "intended" use of AI systems, reflecting the methodology of EU product safety laws. This strategy necessitates identifying the AI system's specific purpose and usage to accurately evaluate the risk level based on its functional role. Within this framework, the Act classifies AI systems into four risk categories: unacceptable, high, limited, and minimal or no risk.

The structure of the AI Act risk categorization is presented by the EU Commission as a pyramid (see EU Commission's diagram below). At the top of this pyramid are prohibited practices, while at the bottom are applications that pose a minimal risk and entail no regulatory obligations. AI systems that pose higher risks are subject to more extensive obligations, both pre- and post-deployment. Limited-risk applications trigger only transparency and disclosure obligations. While the pyramid illustration of the framework established by the AI Act offers clear visualization, its drawback is that it fails to emphasize the overlaps between categories. Specifically, the criteria employed allow for an AI system to be classified under the limited-risk category while simultaneously qualifying as a high-risk system.

FIGURE 16. The AI Act's risk categories



Source: European Commission

[1143] European Commission, *White Paper on Artificial Intelligence: a European approach to excellence and trust*, see *supra* note 1121.

1) Unacceptable risk: Prohibited AI systems

Certain AI applications are banned from the European Union due to their potential to pose unacceptably high risks and their conflict with EU values by infringing on fundamental rights. Article 5 bans—as prohibited "artificial intelligence practices"—the placing on the market, putting into service, or use of AI systems to:

- assess or classify individuals based on their social behavior or their "known, inferred, or predicted personal or personality characteristics" (**social scoring**) when the social score leads to detrimental or unfavorable treatment that is unjustified or disproportionate to the behavior or its gravity, or that is implemented in social contexts unrelated to the contexts in which the data was collected;[1144]

- assess or predict the **risk that a natural person may commit a criminal offense** based solely on the profiling of a natural person or on assessing their personality traits and characteristics, except when AI systems are used to support human assessments based on objective, verifiable facts linked to criminal activity;[1145]

- **exploit vulnerabilities of individuals** "due to their age, disability or a specific social or economic situation" with the objective or effect of distorting the behavior of that person or other persons in a manner that causes or is reasonably likely to cause significant harm;[1146]

- deploy "**subliminal techniques** beyond a person's consciousness or use purposefully manipulative or deceptive techniques" to cause the person to make a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause significant harm to that person or other persons;[1147]

- use **real-time remote biometric identification in publicly accessible spaces** by law enforcement, with only a few exceptions: to search for specific victims and missing persons, prevent serious threats (such as a terrorist attack), or locate or identify a person suspected of having committed a serious criminal offense;[1148]

- use **biometric categorization systems** of individuals based on data that could reveal their race, political views, trade union membership, religious or philosophical beliefs, or sexual orientation, **except in the context of law enforcement**;[1149]

[1144] AI Act, art. 5(1)(c).
[1145] *Id.* art. 5(1)(d).
[1146] *Id.* art. 5(1)(b).
[1147] *Id.* art. 5(1)(a).
[1148] *Id.* art. 5(1)(h).
[1149] *Id.* art. 5(1)(g).

- **infer emotions** of a natural person in the areas of workplace and educational institutions, except for medical or safety purposes (such as monitoring a pilot's fatigue levels);[1150] and

- create or expand **facial recognition databases** through the untargeted scraping of facial images from the internet or CCTV footage.[1151]

2) High-Risk AI systems

AI systems are classified as high risk when their intended purpose presents a high risk of causing harm to the health, safety, or fundamental rights of persons. When an AI system is classified as high risk, it is subject to many requirements and its providers and deployers must comply with stringent obligations.

a) Classification of AI systems as high risk

According to Article 6, AI systems are considered high risk if they fall into one of two categories:

1. The AI system is a safety component or a product **subject to** the EU harmonization legislation listed in Annex I of the Act[1152] and required, as such, to undergo a **third-party conformity assessment**. Annex I lists legislation about toys, medical applications (e.g., AI application in robot-assisted surgery), lifts, cableways installations, motor vehicles, etc.

2. The AI system is used for a **specific purpose listed in Annex III** of the Act.[1153] The list includes AI systems that are used, or intended for use, in areas such as:
    - certain **critical infrastructures**, such as those involving road traffic or the supply of water, gas, and electricity;
    - **education and vocational training**, for purposes such as evaluating learning outcomes (e.g., scoring of exams), guiding the learning process, or determining access to educational institutions;
    - **employment**, workers' management, and access to employment, such as AI tools for recruiting people (e.g., CV-sorting software for recruitment procedures) or to make decisions affecting work relationships;
    - access to **essential private and public services**, including healthcare, evaluation of individuals' creditworthiness, and risk assessment and pricing of life and health insurance;
    - specific applications in **law enforcement**, for example, to assess the risk of a natural person to become a victim of criminal offenses or to evaluate the reliability of evidence;
    - **migration**, asylum, and border control management, for example, to assess the risk of irregular migration posed by a person or examine applications for asylum or verify the authenticity of travel documents;

[1150] *Id.* art. 5(1)(f).
[1151] *Id.* art. 5(1)(e).
[1152] *Id.* art. 6(1).
[1153] *Id.*, art. 6(2).

- **administration of justice**, such as applying the law to a concrete set of facts;
- **democratic processes**, such as influencing the outcome of an election or the voting behavior of persons;
- **non-banned biometrics**, such as remote biometric identification systems and biometric categorization;
- **emotion recognition** systems; and
- **profiling** of natural persons.

Since the high-risk category is particularly broad, the AI Act introduces an exception for AI systems that "do not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making."[1154] This is the case for AI systems intended to:

- perform a narrow procedural task;
- improve the result of a previously completed human activity;
- perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III; and
- detect decision-making patterns or how they diverge from prior patterns of decision-making, as long as the AI system is not meant to replace or influence the previously completed human assessment without proper human review.

The Commission will provide guidelines about the interpretation of the previous lists, including practical examples of high-risk and non-high-risk use cases.[1155] The Commission is also empowered to adopt delegated acts to maintain and update the list of high-risk AI systems[1156] and the list of exceptions.[1157]

b) Requirements applicable to high-risk AI systems

High-risk AI systems must comply with stringent requirements.[1158] These requirements encompass several areas:

- Risk management[1159]
- Data quality and governance[1160]
- Comprehensive technical documentation[1161]
- Consistent recordkeeping[1162]
- Transparency and provision of information to deployers[1163]

[1154] *Id.* art. 6 (3).
[1155] *Id.* art. 6 (5).
[1156] *Id.* art. 7.
[1157] *Id.* art. 6(6).
[1158] *Id.* art. 8 to 15.
[1159] *Id.* art. 9.
[1160] *Id.* art. 10.
[1161] *Id.* art. 11, Annex IV.
[1162] *Id.* art. 12 and 20.
[1163] *Id.* art. 13.

- Guarantee of human oversight[1164]
- Ensuring system accuracy, robustness, and cybersecurity[1165]

A few aspects are worth highlighting:

- **The training, validation, and testing of datasets must be relevant, sufficiently representative, and, to the greatest extent possible, free of errors and complete**, in view of the intended purpose.[1166] This implies protocols for how datasets are created and managed and includes measures for assessing and counteracting potential biases.

- AI systems must be designed to enable **human oversight** while in use.[1167] This implies allowing human operators to detect and understand anomalies and to intervene or disregard the AI system's outputs, particularly to safeguard fundamental rights. Specifically, high-risk AI systems require the imposition of limitations that the system itself cannot bypass and that ensure responsiveness to human operators. Furthermore, they must incorporate mechanisms to guide and inform the designated human supervisor. This guidance is crucial to empowering the supervisor to make well-informed decisions about whether, when, and how to intervene to avert negative consequences or stop the system if it does not perform as intended.

- Lastly, high-risk AI systems "shall be designed and developed in such a way that they achieve **an appropriate level of accuracy, robustness, and cybersecurity**, and that they perform consistently in those respects throughout their lifecycle."[1168] In particular, high-risk AI systems that continue to learn after being placed on the market or put into service must be "developed in such a way as to eliminate or reduce as far as possible the **risk of possibly biased outputs influencing input for future operations"** ('feedback loops') and as to ensure that any such feedback loops are addressed with appropriate mitigation measures.[1169]

c) Obligations applicable to providers of high-risk AI systems

Many obligations are imposed on providers of high-risk AI systems. These obligations are also imposed on third parties that are considered as "providers" under different circumstances.

i. Identification of providers of high-risk AI systems

[1164] *Id.* art. 14.
[1165] *Id.* art. 15.
[1166] *Id.* art. 10.
[1167] *Id.* art. 14.
[1168] *Id.* art. 15(1).
[1169] *Id.* art. 15 (4).

A provider "develops an AI system or a general-purpose AI model" —or "has an AI system or a general-purpose AI model developed" —and "places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge."[1170] The AI Act also considers as "providers" any distributors, importers, deployers, and other third parties that meet certain conditions listed in Article 25(1) (*see figure 17 below*).

In particular, anyone who substantially modifies a high-risk AI system is considered to be a provider.[1171] According to Article 3(23), a "substantial modification" refers to any change that was "not foreseen or planned in the initial conformity assessment." The normal evolution that occurs as part of a machine-learning model's expected development does not count as substantial modifications, as outlined in Recital 128. In addition, anyone who modifies the intended purpose of an AI system, including a general-purpose AI system, and turns it into a high-risk AI system is also considered a provider. This is the case, for example, where a general-purpose AI system is fine-tuned for a particular use that turns out to be high risk.[1172]

FIGURE 17. Qualifications for providers of high-risk AI systems

	These are considered providers of high-risk AI systems
Article 25(1)	Any distributor, importer, deployer, and other third party that: ● puts its **name or trademark** on a high-risk AI system already placed on the market or put into service; ● makes a **substantial modification** to a high-risk AI system in a way that it remains a high-risk AI system; or ● **modifies the intended purpose** of an AI system, including a general-purpose AI system, in a way that the AI system becomes a high-risk AI system.
Article 3(23):	● A **substantial modification** is **a change** to the AI system after it is placed on the market or put into service **which is not foreseen or planned** in the initial conformity assessment by the provider, and as a result the compliance of the AI system (with the requirements set out for high-risk AI systems) is affected or results in a modification to the intended purpose for which the AI system has been assessed.
Recital (128)	● "[W]henever a change occurs which may affect the compliance of a high-risk AI system with the Regulation (e.g., change of operating system or software architecture), or when the intended purpose of the system changes, that AI system should be considered to be a new AI system which should undergo a new conformity assessment." ● "[C]hanges occurring to the algorithm and the performance of AI systems which continue to 'learn' after being placed on the market or put into service, namely automatically adapting how functions are

[1170] *Id.* art. 3(3).
[1171] *Id.* art. 25(1)(b).
[1172] *Id.* art. 25(1)(c).

	carried out, should not constitute a substantial modification, provided that those changes have been pre-determined by the provider and assessed at the moment of the conformity assessment."

The "substantial modifications" criterion is likely to impact small businesses. While major technology firms or well-capitalized startups can afford to create state-of-the-art GPAI models, smaller companies tend to adapt existing advanced models for specific applications. In this scenario, these smaller entities will be subject to the provisions applicable to providers if they make "substantial modifications" and their applications happen to be high risk. The problem is that these small businesses will have to manage risks associated with the original data and the design choices made during the model's development.

ii. Obligations of providers of high-risk AI systems

Providers of high-risk AI systems (including those considered to be providers in accordance with the provisions set out above) are subject to strict obligations.[1173] The obligations presented below are the main obligations the AI Act will impose on providers. Deployers and other parties are also subject to specific obligations that will not be detailed here but are in line with the requirements set out below.

- *Conformity assessments*[1174]

Before introducing a high-risk AI system to the EU market or putting it into use, providers are required to conduct a conformity assessment. This process ensures the AI system meets the mandatory requirements for high-risk AI systems listed above. Conformity assessments for AI systems can be performed by the AI system's provider or by external third-party entities. In most cases, providers are allowed to self-certify conformity.[1175] However, in very limited cases—for biometric systems—a third-party conformity assessment is required to be performed by an accredited independent assessor ("notified body").[1176] When providers conduct their assessments, they must follow strict procedures. In this framework, assessments carried out by independent third parties provide an extra level of scrutiny, which explains why, even without requirements, some companies may still choose to contract with notified bodies for independent evaluations.

After carrying out the conformity assessment (or having it carried out by a notified body), providers must label their AI systems with a CE mark, indicating conformity to EU standards and compliance with the AI Act. Finally, they must register their system on a Commission-managed database for "high risk" AI systems that will be publicly accessible.[1177]

[1173] *Id.* art. 16 to 30.

[1174] *Id.* art. 16(f) and art. 43.

[1175] AI Act, Recital 125: "the conformity assessment of such systems should be carried out as a general rule by the provider under its own responsibility, with the only exception of AI systems intended to be used for biometrics."

[1176] Organizations designated to evaluate and certify high-risk artificial intelligence (AI) systems receive the title of "notified bodies" upon obtaining an official approval notification from a designated government agency, referred to as the "notifying authority."

[1177] *Id.* art. 71.

Market surveillance authorities are empowered to assess systems they suspect have been incorrectly classified and can mandate corrective actions. Furthermore, if a market surveillance authority finds that a provider has improperly classified its AI system to avoid compliance with regulations pertaining to high-risk AI systems, the provider will face fines.

- *Quality management system*[1178]

After the AI system is on the market, providers must create and document a quality management system that ensures compliance with the AI Act. The goal is to reduce the AI system's high risks to an acceptable residual risk level and to implement adequate mitigation and control measures when risks cannot be eliminated. The quality management system shall be documented in a systematic and orderly manner in the form of written policies, procedures, and instructions. It may include, among other elements, a strategy for regulatory compliance, technical specifications, systems and procedures for data management, or the details of a post-market monitoring system.

The AI Act mandates that providers also implement a system for reporting serious incidents as part of their post-market monitoring responsibilities. A serious incident is an event or malfunction that directly or indirectly causes death or leads to serious damage to an individual's health; causes a serious and irreversible disruption of the management and operation of critical infrastructure; causes serious damage to property or the environment; or infringes on fundamental rights under EU law. Both providers and, in certain situations, deployers are required to inform the competent authorities about such incidents. Deployers must inform the provider when they have identified any serious incident. Finally, providers must keep detailed records and the logs automatically generated by their high-risk AI systems.[1179]

- *Detailed documentation*[1180]

To aid in risk management, AI model providers must release detailed documentation of the general characteristics, capabilities and limitations, general logic of the AI system and of the algorithms, system architecture, training methodologies and techniques, training datasets used, and validation and testing procedures used, as well as documentation on the relevant risk management system. The technical documentation should be kept up to date throughout the lifetime of the AI system. Providers must also provide instructions of use with the characteristics, capabilities, and limitations of performance of the AI system.

- *Corrective actions*[1181]

When providers of high-risk artificial intelligence systems detect or have substantial reasons to suspect noncompliance of a system, they must immediately initiate corrective actions. These corrective actions can

[1178] *Id.* art. 17.
[1179] *Id.* art. 19.
[1180] *Id.* art. 11 and Annex IV.
[1181] *Id.* art. 20.

involve modifying the system for compliance, removing it from the market, deactivating it, or initiating a recall, depending on the situation. Additionally, providers are responsible for informing the distributors of the system and, where relevant, the deployers or importers.

When providers of high-risk AI systems detect or suspect that their system presents a risk, they must swiftly initiate an investigation to determine the causes. Additionally, deployers who have reasons to consider that an AI system presents a risk must inform the provider or distributor and relevant market surveillance authority and suspend the use of the system. The provider's investigation should be conducted collaboratively with the deployer who identified the issue, when relevant. Following this, the provider is obligated to report to the market surveillance authorities, explicitly outlining the probe and any corrective measures adopted.

d) Fundamental Rights Impact Assessments (FRIAs) for certain deployers of High-Risk AI systems[1182]

The AI Act requires deployers of certain high-risk AI systems to carry out a "fundamental rights impact assessment" prior to putting the system into use. The deployers targeted by this rule "are bodies governed by public law, or private operators providing public services and operators deploying certain high-risk AI systems" listed in the Act.[1183] This covers banks, insurance companies, and companies active in education, healthcare, or housing.

This fundamental rights impact assessment should include a detailed description of the processes in which the high-risk AI system will be utilized, the duration and frequency of its intended use, the categories of individuals and groups likely to be impacted by its use in the specific context, and the specific risks of harm that could affect these categories or groups. It should also describe the implementation of human oversight measures and the steps to be taken if risks materialize. The impact assessment should apply to the first use of the high-risk AI system and should be updated when the deployer considers that any of the relevant factors have changed.

3) Limited-Risk AI systems requiring transparency[1184]

For some AI systems, a critical requirement is transparency, especially in cases where there is a significant risk the system can or will be used to deceive or manipulate people without their knowledge or consent. Four categories of tools are listed in Article 50: chatbots, applications designed to create deepfakes, generative AI tools, and tools designed to recognize emotions or categorize biometrics. These AI systems that are subject to transparency obligations are often presented as "limited risk" (or "specific transparency risk") applications.[1185] However, some of them, such as emotion recognition systems, are also high risk if

[1182] *Id.* art. 27.
[1183] *Id.* art. 27.
[1184] *Id.* art. 50.
[1185] European Commission, *Why do we need to regulate the use of Artificial Intelligence?,* Artificial Intelligence – Q&As, EUROPEAN COMMISSION (Dec. 12, 2023), https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683

they fit in the lists mentioned above (*see section 5.1.2.B.2*). In such cases, the obligations applicable to high-risk AI systems will also apply. Additionally, these systems subjected to transparency obligations can also belong to the category of general-purpose AI systems (*see section 5.1.2.C*) and, if so, they will be subject to the corresponding obligations.

i) Chatbots[1186]

Providers of chatbots must make sure that end users are aware they are speaking with a machine. The responsibility to disclose that users are interacting with AI rather than humans falls to the providers, not deployers or users.

FIGURE 18. Transparency obligation for chatbots

Article 50(1) applicable to	**Obligations**
Providers of AI systems (such as chatbots) intended to interact directly with natural persons	• The system must be **designed and developed in such a way that the concerned natural persons are informed that they are interacting with an AI system** • unless this is obvious from the point of view of a natural person who is reasonably well-informed, observant, and circumspect, considering the circumstances and context of use. • Exception: AI systems **authorized by law to detect, prevent, investigate, and prosecute criminal offenses**, unless those systems are available for the public to report a criminal offense.

ii) Deepfakes applications[1187]

In the case of deepfake technologies, it is the responsibility of the deployers (i.e., anyone using an AI system under their authority) to inform third parties about their use.

FIGURE 19. Disclosure obligations for deepfakes

Article 50(4) applicable to	**Obligations**
Deployers of AI systems generating deepfakes (AI systems that generate	

[1186] AI Act, art. 50(1).
[1187] *Id.* art. 50(4).

or manipulate images, audio, or video constituting a "deepfake")	• Deployers shall **disclose that the content has been artificially generated or manipulated**. • Exceptions: • Where the use **is authorized by law to detect, prevent, investigate, and prosecute criminal offenses**. • Where the content forms part of an **evidently artistic, creative, satirical, or fictional analogous work or program.** In such cases, the transparency obligations are limited: The deployer must disclose the existence of generated or manipulated content in an appropriate manner that does not hamper the display or enjoyment of the work.

iii) Generative AI systems

Two provisions of Article 50 deal with generative AI systems. Article 50(2) is aimed at AI systems generating synthetic audio, image, video, or text content. Article 50(4) is aimed at AI systems that generate or manipulate text published to inform the public on matters of public interest. The latter provision clearly targets the media and public relations industry and the way they may use generative AI tools to create press releases, news articles, and blog posts. Article 50(4) covers not only the generation of content but also the "manipulation" of existing content, which harkens to the issue of deepfakes, mentioned above.

For generative AI systems, providers must ensure that the content is marked as machine-generated, using such means as watermarking.[1188] Deployers of AI systems used for generating or manipulating text are obligated to inform the public that the text was artificially generated, except in cases where the content was reviewed by a natural or legal person who has editorial responsibility for the text.

FIGURE 20. Disclosure and marking obligations for AI generated content

Provisions applicable to	**Obligations**
Article 50(2) **Providers** of AI systems, including general-purpose AI systems, generating synthetic audio, image, video, or text content (i.e. generative AI systems)	• Providers shall ensure the outputs of the AI system **are marked in a machine-readable format and detectable as artificially generated or manipulated** (watermarking). • Technical solutions must be effective, interoperable, robust, and reliable as far as this is technically feasible, considering specificities and limitations of

[1188] *See* above section 4.1.3.C.(2).

	different types of content, costs of implementation, and the generally acknowledged state-of-the-art, as may be reflected in relevant technical standards. ● Exceptions: - when the AI systems perform an assistive function for standard editing; - when the AI systems do not substantially alter the input data provided by the deployer or the semantics thereof; or - when authorized by law to detect, prevent, investigate, and prosecute criminal offenses.
Article 50(4) **Deployers** of AI systems that generate or manipulate text published to inform the public on matters of public interest	● Deployers shall disclose that the text has been artificially generated or manipulated. ● Exceptions: ● Where the use is authorized by law to detect, prevent, investigate, and prosecute criminal offenses, or ● When the AI-generated content has undergone a process of human review or editorial control, and where a natural or legal person holds editorial responsibility for the publication of the content.

iv) Emotion recognition systems or biometric categorization systems[1189]

Article 50(3) of the EU's AI Act requires that deployers shall inform the natural persons exposed to a deployer's emotion recognition or biometric categorization system that they have been or are exposed to the system. This covers, for instance, Facial Emotion Recognition (FER), "the technology that analyzes facial expressions from both static images and videos in order to reveal information on one's emotional state."[1190] Article 50(3) also covers emotion recognition systems processing and analyzing data posted on social media, in order to infer emotions.[1191] Biometric categorization systems are AI systems designed to assign individuals to specific categories based on their biometric data, including characteristics such as sex, age, hair color, eye color, tattoos, ethnic background, or sexual or political orientation.

FIGURE 21. Disclosure obligation for the use of emotion recognition and biometric systems

Article 50(3) applicable to	**Obligations**

[1189] AI Act, art. 50(3).

[1190] European Data Protection Supervisor, *Facial Emotion Recognition*, TechDispatch, no.1, 2021, at 1.

[1191] *See* Chen Li, Fanfan Li, *Emotion Recognition of Social Media Users Based on Deep Learning*, PeerJ Computer Science (2023) 9:e1414, https://doi.org/10.7717/peerj-cs.1414; Luis Romero Gomez et al., *Emotion Recognition on Social Media Using Natural Language Processing (NLP) Techniques*, Proceedings International Conference on Information Science & Systems, 113 (Nov. 21, 2023) https://doi.org/10.1145/3625156.3625173.

Deployers of an emotion recognition system or a biometric categorization system	• Shall inform the natural persons exposed to the system • Exception: for systems permitted by law to detect, prevent, and investigate criminal offenses

4) Minimal-risk AI systems

All other AI systems, not targeted by the AI Act, can be developed and utilized in accordance with other existing legal frameworks and do not incur any extra obligations. According to the EU Commission, most AI systems currently in use in the EU belong to this category.[1192] Examples include applications such as AI-enabled recommender systems or spam filters. Providers of these systems can voluntarily adhere to the standards for trustworthy AI and are encouraged to follow voluntary codes of conduct.

5.1.2.C General-Purpose AI (GPAI) models

Although the initial draft of the AI Act by the European Commission did not specifically address generative AI, this did not mean that these models were necessarily outside the scope of the draft Regulation. All AI systems can, in fact, fall into the various risk categories presented in the previous section. However, the risk classification of the AI Act is based on the intended purpose of these AI systems. This means that a foundation model that has not yet been fine-tuned for a specific use case cannot be precisely regulated by these provisions. In practice, foundation models often lack a predefined purpose and demonstrate exceptional versatility and learning capabilities, enabling them to undertake new tasks in often unpredictable ways.

In this context, it would have been tempting for the EU drafters to include foundation models in the category of high-risk applications. This would have meant abandoning a classification based entirely on industry sectors and use cases. In September 2022, OpenAI transmitted to the EU Commission and Council a white paper arguing that, although OpenAI's GPT-3 model could be deployed in high-risk applications, it should not be automatically classified as a high-risk system.[1193] Meanwhile, critics of the AI Act draft pointed out that the provisions were excessively rigid and not adaptable to the current state of AI technology.[1194] It became evident that foundation models would necessitate regulations tailored to their specific characteristics, despite uncertainties about their potential uses and future risks.

Therefore, during negotiations, the European drafters introduced new provisions on foundation and generative AI models. The EU Council created a new category of "general-purpose AI systems." These were defined by the Council as any AI system that "is intended by the provider to perform generally

---

[1192] European Commission, *AI Act: Shaping Europe's Future*, European Commission, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (last visited Apr. 6, 2024).

[1193] Billy Perrigo, *Exclusive: OpenAI Lobbied the E.U. to Water Down AI Regulation*, Time (June 20, 2023), https://time.com/6288245/openai-eu-lobbying-ai-act/.

[1194] Kai Zenner, *A Law for Foundation Models: The EU AI Act Can Improve Regulation for Fairer Competition*, OECD.AI Policy Observatory (July 20, 2023), https://oecd.ai/en/wonk/foundation-models-eu-ai-act-fairer-competition.

applicable functions" and that may be used in a "plurality of contexts and be integrated into a plurality of other AI systems."[1195] The EU Council's version of the AI Act specified that, when they *may* be used as high-risk AI systems, GPAI systems must comply with regulatory requirements that closely mirror those applicable to high-risk AI systems.

A few months later, the EU Parliament added the concept of "foundation model" in its own version of the AI Act.[1196] The Parliament's draft provided for a number of obligations on providers of foundation models. Among other things, they would be obliged to create technical documentation, establish data governance measures to assess the suitability of datasets, establish a quality management system, and register the model in the EU database. Additionally, the Parliament's draft included specific provisions for generative AI *systems*.[1197]

During the Trilogues (meetings of representatives from the European Parliament, Commission, and Council), the negotiators agreed to regulate the most dangerous models more closely, because of their capacity and reach. The Spanish Presidency of the EU proposed stricter regulation of "very capable foundation models" and "general purpose AI systems built on foundations models and used at scale in the EU."[1198]

In the end, the final version of the Regulation specifically targets GPAI *models*, and not GPAI *systems* as in the Council's draft. The AI Act includes a tiered approach that distinguishes between GPAI models and GPAI models with "systemic risk," imposing stricter obligations on the latter. This tiered approach results from a compromise among some Member States (such as France, Germany, and Italy) who opposed any regulation of foundation models, and the European Parliament, which favored imposing uniform obligations on all foundation models. European authorities have maintained the pyramid illustration for the framework, and they have expanded its base to include the category of GPAI models, upon which GPAI systems can be built.

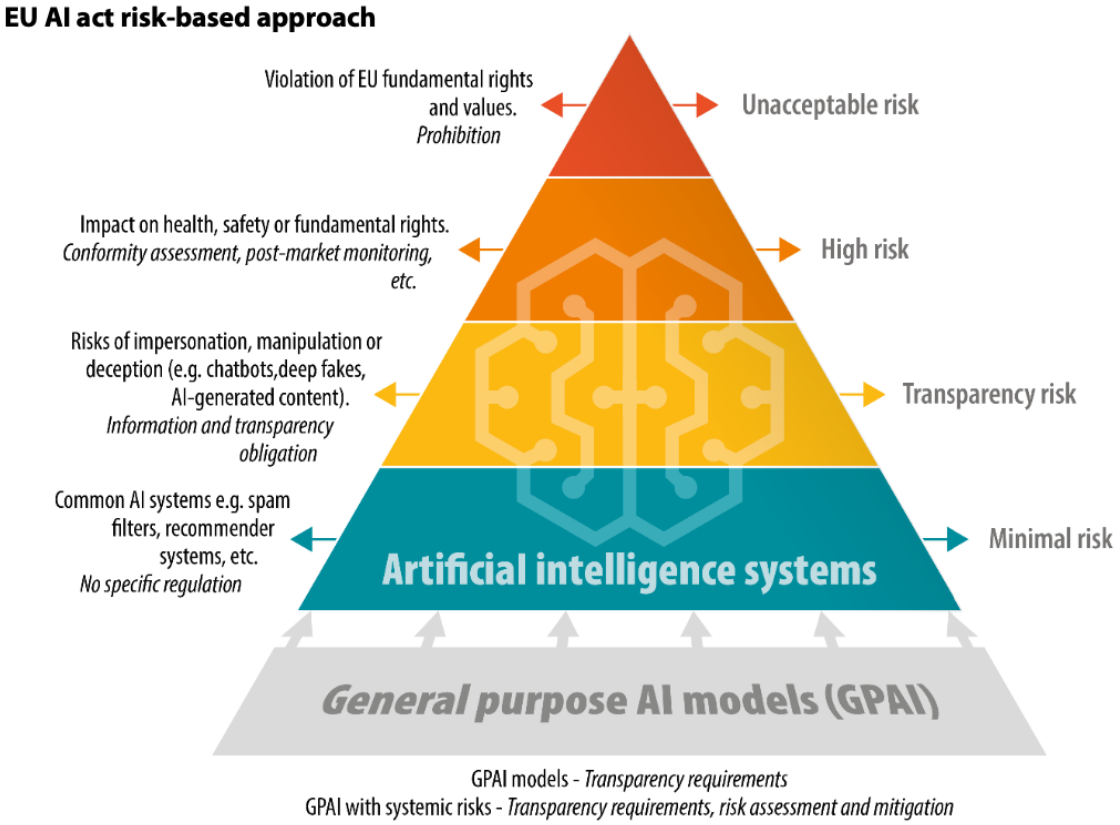
[1195] art. 3(1b) of the EU Council's Common Position (General Approach), December 6, 2022 (*see Appendix II)*; Council of the European Union, Proposal for a regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) - General Approach, 2021/0106(COD), https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf.

[1196] EU Parliament, *Amendments on the proposal for an AI Act* (June 14, 2023), *see Appendix III;* Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act), P9_TA(2023)0236, https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

[1197] *Id.*

[1198] Proposal of the Spanish Presidency during the Trilogues, see *Appendix IV*; Luca Bertuzzi, *AI Act: EU countries headed to tiered approach on foundation models amid broader compromise,* EURACTIV (October 17, 2023) https://www.euractiv.com/section/artificial-intelligence/news/spanish-presidency-pitches-obligations-for-foundation-models-in-eus-ai-law/.

FIGURE 22. The AI Act pyramid



Data source: European Commission

1) General-Purpose AI Models

The finalized version of the AI Act differentiates between general-purpose AI *models* and *systems*. This distinction is established by Articles 3(63) and 3(66). GPAI models do not constitute AI systems on their own: they are essential components of AI systems and require further components to become AI systems. When GPAI models, as described in the Act (*see Figure 23 below*), are identified, the provisions of Chapter 5 become applicable.[1199] The fact that these provisions apply specifically to GPAI *models* and not to GPAI *systems* exemplifies that, in the specific context of general-purpose AI, the provisions of the Act govern the technology itself rather than its applications.

FIGURE 23. Definitions of GPAI model and GPAI system

	Definitions
GPAI model *Article 3(63)*	"General-purpose AI model" means an AI model, including when trained with **a large amount of data using self-supervision at scale**, that displays **significant generality** and is capable of competently performing **a wide range**

[1199] AI Act, art. 51 to 56.

	of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems of applications, except AI models that are used for research, development, and prototyping activities before they are placed on the market.
GPAI system *Article 3(66)*	"General-purpose AI system" means **an AI system that is based on a general-purpose AI model** that has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems.
Recital (97)	"Although AI models are essential components of AI systems, they do not constitute AI systems on their own. AI models require the addition of further components, such as for example a user interface, to become AI systems. AI models are typically integrated into and form part of AI systems."

The AI Act requires GPAI models to meet specific requirements.

a) Technical documentation

Providers of general-purpose AI models must produce technical documentation that includes information about the way the model was designed and developed and the main characteristics of the model. In particular, the AI Act requires the disclosure of technical details about the way the model was trained—for example, about the training data or the computational resources used to train the model.[1200]

It is also worth highlighting that the "known or estimated" energy consumption of the model must be disclosed. The text does not say whether it is the energy consumed to *train* the model or to *operate* it. It is also specified that, if the energy consumption is not known, it is sufficient to disclose the computational resources used. Overall, these requirements remain relatively vague.

FIGURE 24. Technical documentation about GPAI models

	Obligation to draw up technical documentation
Article 53(1)(a)	• Providers of GPAI models must draw up and keep up-to-date the **technical documentation** of the model for the purpose of providing it, upon request, to the AI Office and the national competent authorities. • The technical documentation must include information about the training and testing process of the model, the results of its evaluation, and the information listed in Annex XI (below) as appropriate to the size and risk profile of the model.
Annex XI **Section 1(1)**	The technical documentation must include a **general description of the GPAI model** including:

[1200] AI Act, Annex XI.

	● the tasks that the model is intended to perform and the type and nature of AI systems in which it can be integrated; ● the acceptable use policies applicable; ● the date of release and methods of distribution; ● the architecture and number of parameters; ● the modality (e.g., text, image) and format of inputs and outputs; and ● the license.
Annex XI **Section 1(2)**	The technical documentation must include ● a **detailed description** of the elements of the GPAI model referred to in Section 1(1) above ● and relevant information about the **process used for its development**, including: ○ the **technical means** (e.g., instructions of use, infrastructure, tools) required for the GPAI model to be integrated in AI systems; ○ the **design specifications** of the model and training process, including ■ training methodologies and techniques; ■ the key design choices, including the rationale and assumptions made; ■ what the model is designed to optimize for; and ■ the relevance of the different parameters, as applicable; ○ information on **the data used** for training, testing, and validation, including: ■ type and provenance of data and curation methodologies (e.g., cleaning, filtering, etc.); ■ the number of data points; ■ their scope and main characteristics; ■ how the data was obtained and selected; and ■ other measures to detect the unsuitability of data sources and methods to detect identifiable biases; ○ the **computational resources used** to train the model (e.g., number of floating-point operations – FLOPs), training time, and other relevant details related to the training; ○ known or estimated **energy consumption** of the model; in case it is not known, this could be based on information about computational resources used.

b) Obligation to provide information and documentation to downstream providers

General-purpose AI systems may be used as systems by themselves or serve as components of other AI systems. Therefore, due to their particular nature and to ensure a fair sharing of responsibilities along the AI supply chain, the providers of such systems should closely cooperate with the providers of the respective

high-risk AI systems to enable their compliance with the relevant obligations. Providers of general-purpose AI models must, therefore, provide sufficient information to downstream providers so that they can use the model or even fine-tune the model appropriately.

FIGURE 25. Documentation on GPAI models to be provided to downstream providers

	Obligation to provide information and documentation to downstream providers
Article 53(1)(b)	Providers of GPAI models must draw up, keep up-to-date, and make available **information and documentation to providers** who intend to integrate the GPAI model into their own AI system.
Article 53(1)(b)	The information and documentation shall: (1) enable providers of AI systems to have **a good understanding of the capabilities and limitations of the GPAI model and to comply** with their own obligations pursuant to the Regulation; and (2) contain, at a minimum, the elements listed in Annex XII (below).
Annex XII (1)	The information and documentation for other providers shall include **a general description** of the GPAI model including: ● the tasks that the model is intended to perform and the type and nature of AI systems into which it can be integrated; ● the acceptable use policies applicable; ● the date of release and methods of distribution; ● how the model interacts or can be used to interact with hardware or software that is not part of the model itself, where applicable; ● the versions of relevant software related to the use of the GPAI model, where applicable; ● the architecture and number of parameters, ● the modality (e.g., text, image) and format of inputs and outputs; and ● the license for the model.
Annex XII (2)	The information and documentation for other providers shall include **a description of the elements of the model and of the process for its development**, including: ● the technical means (e.g., instructions of use, infrastructure, tools) required for the GPAI model to be integrated into AI systems; ● modality (e.g., text, image, etc.) and format of the inputs and outputs and their maximum size (e.g., context window length, etc.); and ● information on the data used for training, testing, and validation, where applicable, including type and provenance of data and curation methodologies.

Article 53(7)	Any information or documentation obtained pursuant to Article 53, including trade secrets, shall be treated in compliance with the confidentiality obligations set out in Article 78, in particular "the intellectual property rights and confidential business information or trade secrets of a natural or legal person, including source code" (Article 78(1)(a)).

c) Copyright policy

Interestingly, the drafters of the AI Act have included provisions designed to force GPAI model developers to comply with the EU copyright law (*see above section 5.1.1B*). They must, therefore, put in place a company policy designed to ensure that their models are not trained on data collected illegally or on data collected in violation of the reservations expressed by copyright holders in the forms provided for by Article 4(3) of the New Copyright Directive.[1201] In scraping the web for training data, AI companies are obliged to put in place the technical means necessary to enable rights holders to opt out and express it in "an appropriate manner" according to the Text and Data Mining exception of the New Copyright Directive.

FIGURE 26. Copyright policy of providers of GPAI models

	Obligation to put in place a policy to comply with EU copyright law
Article 53(1)(c)	Providers of GPAI models should put in place a policy: • to comply with European Union law on copyright and related rights • in particular to identify and provide, including through state-of-the-art technologies, a reservation of rights expressed by rights holders pursuant to the Text and Data Mining exception (Article 4(3) of Directive (EU) 2019/790, see below)
Article 4(3) of Directive (EU) 2019/790	Web scraping is permitted only on condition that the use of protected materials "has not been expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online." (Text and Data Mining exception)
Recital (106)	Any provider placing a general-purpose AI model on the Union market should comply with this obligation, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place.

[1201] AI Act, art. 4(3).

	This is necessary to ensure a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union.

Here, the AI Act not only reiterates existing copyright rules; it also appears to attempt to expand the territorial scope of the existing EU's Text and Data Mining (TDM) rules, given the extraterritorial scope of the Act.[1202] Recital 106 explicitly provides that every provider of GPAI models must adhere to the requirements of the Text and Data Mining exception, particularly concerning the rights holders' ability to reserve the use of their work. This obligation is imposed on any providers putting a GPAI model in the EU market "regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place," as provided by Recital 106. The impact of this provision may be significant, particularly because training datasets are often created far from where the models are ultimately used. Recital 106 implies that AI developers who offer models trained *outside* the EU, such as in the US, must demonstrate that they have enabled the authors of the content used in the training to reserve their rights, even if these authors are not based in the EU.

Nevertheless, such an interpretation warrants further discussion. This broad expansion of the territorial scope of EU copyright rules stems from Recital 106 of the AI Act. However, recitals primarily aim to elucidate the rules set forth in EU legislation, not establish binding principles. Recital 106 seems insufficient to broaden the territorial scope of EU copyright law significantly. For the time being, the TDM exception, as outlined in the New Copyright Directive, should cover only those training activities with a connection to EU territory, such as when developers scrape data from European websites.

d) Transparency on training datasets

Developers of GPAI models will have to provide detailed summaries of the data they scrape. However, this requirement does not necessarily mean that these summaries provide extensive details, such as the list of the specific websites where data were scraped. It appears sufficient to list the main databases used, for example. In any case, the requirement to publish general descriptions should adequately enable third parties to verify whether model providers have trained their models on legally accessible data sources, as required by the Text and Data Mining exception.[1203]

Furthermore, there might be some overlap regarding training data transparency with the distinct obligation to prepare and maintain the technical documentation that "must include information about the training and testing process of the model."[1204] However, the requirement for technical documentation does not extend to open-source GPAI models: Providers of open-source models will thus be obliged to provide information about their training datasets.

[1202] Keller, *see supra* note 1066.
[1203] *See* New Copyright Directive 2019/790, art. 4(1).
[1204] AI Act, art. 53(1).

FIGURE 27. Transparency about training datasets

	Obligation to disclose content used for training
Article 53(1)(d)	Providers of GPAI models must draw up and make publicly available a **sufficiently detailed summary about the content used for training** of the GPAI model, according to a template provided by the AI Office.
Recital (107)	This summary should: ● Take into due account the need to protect trade secrets and confidential business information, and ● be **generally comprehensive in its scope, instead of technically detailed, to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights** under Union law. For example, it could: ● list the main data collections or sets that went into training the model, such as large private or public databases or data archives, or ● and provide a narrative explanation about other data sources used. AI Office will provide a template for the summary that should be simple and effective, and allow the provider to provide the required summary in narrative form.

e) Exemption for open-source models

Even though the AI Act provides that it does not cover free and open-source systems and models, this exclusion is notably restricted.[1205] The AI Act regulates free and open-source prohibited models, high-risk models, models requiring transparency obligations, or GPAI models. However, providers of free and open-source GPAI models are exempt from the obligation to release technical documentation and provide detailed information to downstream providers.

[1205] AI Act, art. 2(12).

The criteria to qualify for this exemption under the Act are stringent. To qualify, AI models must be entirely open, entailing the full disclosure of all parameters and allowing for their unrestricted use, including modifications. This openness is the rationale for exempting such models from transparency obligations. Furthermore, to be eligible for the exemption, the distribution of these models must occur free of charge, which rules out any indirect monetization. Compliance with these conditions results in the waiver of specific obligations, namely the provision of technical documentation and information to downstream users. Nonetheless, adherence to all other requirements, such as copyright policy and the summary of training data, remains mandatory.

FIGURE 28. Open-source exemption

	Exemption for providers of free and open-source models
Article 53(2)	The obligations to draw up technical documentation and provide information to downstream developers do not apply to providers of AI models that are released under a free and open license: – that allows for the access, usage, modification, and distribution of the model, and – whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available. This exception does not apply to GPAI models with systemic risks.
Recital (103)	● Free and open-source AI components: – cover software and data, including models and GPAI models, tools, services, or processes of an AI system, – can be provided through different channels, including their development on open repositories. ● AI components are not free and open source when they are provided against a price or otherwise monetized, including: – through the provision of technical support or other services, including through a software platform related to the AI component – through the use of personal data for reasons other than exclusively for improving the security, compatibility and interoperability of the software with the exception of transactions between microenterprises ● the fact of making AI components available through open repositories should not, in itself, constitute a monetization

2) General-Purpose AI models with "systemic risk"

The concept of applying stricter and more detailed obligations to the riskiest AI models generated considerable discussion during the AI Act negotiations. The most elaborate proposal in this respect came from the Spanish Presidency of the EU Council during the Trilogues (*See Appendix IV*). The Spanish Presidency proposed stricter regulation of "very capable foundation models" and "general purpose AI systems built on foundation models and used at scale in the EU."[1206] The proposal defined "very capable foundation models" as models "whose capabilities go beyond the current state-of-the-art and may not yet be fully understood." And the proposal suggested the computational power used during training should be the primary criterion to determine capability. In the absence of tools and methodologies for predicting and measuring the capabilities of such advanced models, the computational power would be measured in FLOPs.[1207] The Spanish Presidency proposal also provided for specific rules for "GPAI systems built on foundation models and used at scale in the EU." The suggested criteria to identify such systems were related to reach and impact, for example, [10,000] registered business users (i.e., developers) or [45 million] registered end users.

The final version of the AI Act retains the notion of GPAI models with systemic risk, which means a "systemic risk at Union level." Such models not only have "high impact capabilities" but also have a "significant impact on the internal market" and "negative effects" that can be "propagated at scale across the value chain."[1208] The Recitals highlight that "systemic risks should be understood to increase with model capabilities and model reach."[1209]

FIGURE 29. Definition of systemic risk

	Definitions
Systemic risk Article 3(65)	"Systemic risk" means "a risk that is **specific to the high-impact capabilitie**s of GPAI models having **a significant impact on the Union market due to their reach** or due to actual or reasonably foreseeable **negative effects on public health, safety, public security, fundamental rights, or the society as a whole**, that can be **propagated at scale across the value chain**."
High-impact capabilities in GPAI models Article 3(64)	"High-impact capabilities" means "capabilities that match or exceed the capabilities recorded in the most advanced GPAI models."

[1206] Bertuzzi, *supra* note 1198.
[1207] Floating-point operations per second (FLOPS) is a metric used to measure a computer's performance. It quantifies the number of floating-point arithmetic calculations the processor can execute in one second. *Id.*
[1208] AI Act, art. 3(65).
[1209] *Id.* Recital 110.

Systemic risks of GPAI models Recital (110)	● Systemic risks raised by GPAI models include, but are not limited to, any actual or reasonably foreseeable negative effects - in relation to major accidents, disruptions of critical sectors, and serious consequences to public health and safety; - on democratic processes, public and economic security; the dissemination of illegal, false, or discriminatory content. ● Systemic risks: - should be understood to increase with model **capabilities** and model **reach**; - can arise along the entire **life cycle** of the model; and - are influenced by conditions of misuse, model reliability, model fairness and model security, the level of autonomy of the model, its access to tools, novel or combined modalities, release and distribution strategies, the potential to remove guardrails, and other factors. ● International approaches have identified the need to pay attention to: - risks from potential intentional **misuse** or unintended issues of control relating to alignment with human intent; - **chemical, biological, radiological, and nuclear risks**, such as the ways barriers to entry can be lowered, including for weapons development, design acquisition, or use; - offensive **cyber capabilities**, such as the ways vulnerability discovery, exploitation, or operational use can be enabled; - the effects of **interaction** and tool use, including, for example, the capacity to control physical systems and interfere with critical infrastructure; - risks from models of making copies of themselves or '**self-replicating**' or training other models; - the ways models can give rise to **harmful bias and discrimination** with risks to individuals, communities or societies; - the facilitation of **disinformation** or harming **privacy** with threats to democratic values and human rights; and - risk that a particular event could lead to a **chain reaction** with considerable negative effects that could affect up to an entire city, an entire domain activity, or an entire community.

a) Classification of GPAI model with systemic risk

A GPAI model is said to carry a "systemic risk" if:

1. It has high-impact capabilities, which is presumed if it is trained using computing power exceeding 10^{25} FLOPs,[1210] indicative of its capabilities and the volume of data involved.

[1210] This is a lower threshold than the 10^26 FLOPS threshold for the reporting obligation under the U.S. Executive Order on AI (*see below section 5.3.2.B.3*).

2. It is designated as such by the EU Commission, ex officio, or following a qualified alert by the scientific panel, based on specific criteria listed in Annex XIII (*see Figure 30 below*). The Commission can reassess this designation upon request from the provider.

The European Commission will maintain a list of GPAI models with systemic risk.

FIGURE 30. Classification of GPAI models with systemic risk.

	Classification of GPAI models with systemic risk
Article 51(1) and 51(2)	A GPAI model with systemic risk meets any of the following conditions: a. It has **high-impact capabilities**: evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks. ⇒ *A GPAI model is presumed to have high-impact capabilities when the cumulative amount of compute used for its training, measured in floating point operations (FLOPs), is greater than* 10^{25} . b. It is designated as having high-impact capabilities by **a decision of the Commission,** ex officio, or following a qualified alert by the scientific panel, based on the criteria listed in Annex XIII (see below).
Article 3 (67)	"Floating-point operation" means any mathematical operation or assignment involving floating-point numbers, which are a subset of the real numbers typically represented on computers by an integer of fixed precision scaled by an integer exponent of a fixed base.
Annex XIII	The designation of GPAI with systemic risk is made by the Commission based on the following criteria: ● number of **parameters** of the model; ● quality or size of the **dataset**, for example, measured through tokens; ● the **amount of computation** used for training the model, measured in FLOPs or indicated by a combination of other variables, such as estimated cost of training, estimated time required for the training, or estimated energy consumption for the training; ● **input and output modalities** of the model, such as text to text (large language models), text to image, multimodality, and the state-of-the-art thresholds for determining high-impact capabilities for each modality, and the specific type of inputs and outputs (e.g., biological sequences); ● benchmarks and evaluations of capabilities of the model, including the number of tasks without additional training; adaptability to learn new, distinct tasks; degree of autonomy and scalability; and the tools it has access to; ● a high impact on the internal market due to its reach, which shall be presumed when it has been made available to at least 10,000 registered business users established in the Union; and ● number of register**ed end users**.

Articles 51(3) and 52(4)	The Commission is empowered to adopt delegated acts to revise the criteria used to presume or decide that a model has high-impact capabilities.
Article 52(6)	The Commission publishes a list of GPAI models with systemic risk and keeps it up to date.

Providers whose GPAI models meet the criterion of 10^25 FLOPS that leads to the presumption of high-impact capability must notify the Commission.[1211] This threshold of 10^25 FLOPs to presume that a GPAI model carries systemic risk is relatively high. Only the most recent models, such as GPT-4 or Gemini, appear to meet the threshold.[1212] Some scholars have suggested lowering the threshold to 10^24 FLOPs to include other models.[1213] Conversely, some Member States, such as France, have expressed their willingness to raise the threshold.[1214] It is likely, in any case, that the Commission will modify this threshold in a delegated act.

Providers who do not want their model to be classified as systemic risk should be able to demonstrate that, because of its specific characteristics, their model does not present such risk.[1215] Recital 112 states that this information "is especially important with regard to general purpose AI models that are planned to be released as open source." Indeed, there is no exemption for open-source GPAI models with systemic risk: open-source models are subject to all the applicable provisions. However, these provisions will be more difficult to enforce within the context of open source.

b) Obligations of providers of foundation models with systemic risk

The AI Act imposes supplementary obligations on providers of foundation models with systemic risk, in addition to the above-mentioned requirements applicable to all GPAI models. Free and open-source GPAI models with systemic risk do not benefit from any exemption.[1216] Providers of such models must comply with all the requirements applicable to GPAI models and the additional obligations imposed on GPAI models with systemic risk.

[1211] *Id.* art. 52 (1).

[1212] *EU AI Act Compliance Analysis*, THE FUTURE SOCIETY, (Dec. 2023) https://thefuturesociety.org/wp-content/uploads/2023/12/EU-AI-Act-Compliance-Analysis.pdf.

[1213] Philipp Hacker, *Comments on the Final Trilogue Version of the AI Act* (Jan. 23, 2024), THE EUROPEAN NEW SCHOOL OF DIGITAL STUDIES, https://www.europeannewschool.eu/images/chairs/hacker/Comments%20on%20the%20AI%20Act.pdf.

[1214] Alexandre Piquard, *France agrees to ratify the EU Artificial Intelligence Act after seven months of resistance*, LE MONDE (Feb. 3, 2024), https://www.lemonde.fr/en/economy/article/2024/02/03/france-agrees-to-ratify-the-eu-artificial-intelligence-act-after-seven-months-of-opposition_6489701_19.html, *see also* Luca Bertuzzi, *EU countries give crucial nod to first-of-a-kind Artificial Intelligence Law*, EURACTIV (Feb. 2 2024), https://www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/.

[1215] AI Act, art. 52(2).

[1216] *Id.* art. 53(2).

FIGURE 31 Obligations of providers of GPAI models with systemic risk.

	Additional obligations for providers of GPAI models with systemic risk
Article 55	Providers of GPAI models with systemic risk must: ● perform model evaluation in accordance with standardized protocols and tools reflecting the state-of-the-art, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risk; ● assess and mitigate possible systemic risk; ● keep track of, document, and report to the AI Office and national competent authorities, relevant information about serious incidents and possible corrective measures to address them; and ● ensure an adequate level of cybersecurity protection for the model and the physical infrastructure of the model.

It is noteworthy that the AI Act clearly expresses, in Article 55, the obligation to carry out model evaluations, including red teaming (*see above section 4.1.1.B(2)*). This means enshrining in the Act a practice that is widespread in the industry and deemed to be effective. Recital 114 specifies that such adversarial testing can be conducted internally or externally and should take place prior to placing the model on the market.

Additionally, systemic risks must be continuously assessed and mitigated. Recital 110 mentions the risks of major accidents; disruptions of critical sectors and serious consequences to public health and safety; negative effects on democratic processes, and public and economic security; or dissemination of illegal, false, or discriminatory content. Addressing these risks implies implementing a comprehensive risk management system, which may include accountability and governance processes, post-market monitoring, appropriate measures along the entire model's life cycle, and cooperation with relevant actors across the AI supply chain, as outlined by Recital 114.

Finally, providers must clearly detail evaluation and testing practices in the additional information they must disclose within the technical documentation they supply.

FIGURE 32. Additional information in the technical documentation

	Obligation to provide additional information in the technical documentation
Annex XI **Section 2**	Providers of GPAI models with systemic risk must provide additional information in their technical documentation: ● detailed description of the **evaluation strategies**, including evaluation results, criteria, metrics, and the methodology on the identification of limitations. ● detailed description of the measures put in place to conduct: – **internal and/or external adversarial testing** (e.g., red-teaming);

	– and **model adaptations**, including alignment and fine-tuning; and ● detailed description of the **system architecture**, explaining how software components build or feed into each other and integrate into the overall processing.

In summary, the drafters of the AI Act opted to adopt the term "systemic risk" from the Digital Services Act[1217] rather than incorporating commonly used terms such as "frontier models" or "highly capable" models into the legislation. However, the term "systemic risk" is somewhat ambiguous. The criteria for identifying models that pose a systemic risk are also subject to debate, although the Commission and the AI Office will certainly refine these criteria over time. Importantly, the computational resources used to train a model do not necessarily indicate that this model poses a greater risk than a model that was trained with fewer resources. Finally, although the provisions of the AI Act aim to enforce an effective risk mitigation strategy for the targeted companies, similar to the DSA, they lack the guarantees of transparency provided by the DSA. Notably, the AI Act lacks an equivalent to Article 40 of the DSA, which grants vetted researchers access to company data to evaluate the risks posed by their activities and the efficacy of their risk mitigation measures.

5.1.2.D. Other provisions of the AI Act

Other provisions of the AI Act are worth mentioning in this section: the sandbox and real-world testing mechanisms, and the right to explanation of individual decision-making.

1) Sandboxes and real-world testing

The AI Act includes provisions that allow authorities in national Member States to set up "regulatory sandboxes."[1218] Sandboxes encourage innovation in AI by allowing developers to experiment with AI technologies without the full burden of regulatory constraints. These sandboxes operate under the oversight of competent authorities. Under the AI Act framework, a specific sandbox plan is agreed upon between the prospective providers and the competent authority. The sandbox is designed to facilitate the development, training, testing, and validation of innovative AI systems for a limited time before they are placed on the market or put into service. It may include testing in real-world conditions supervised in the sandbox.

The AI Act also encourages real-world testing outside regulatory sandboxes under certain conditions.[1219] This approach allows for experimentation with high-risk AI systems for a limited duration, i.e., a maximum period of six months, extendable by an additional six months. Before commencing testing, the provider or prospective provider must formulate a "real-world testing plan" and submit it for approval to the competent market surveillance authority. This testing is contingent on several safeguards: The subjects of the testing must provide informed consent; the testing should not adversely affect them; the predictions,

[1217] DSA, art. 34; *see also* G'sell, *supra* note 1094.
[1218] AI Act, art. 57.
[1219] *Id.* art. 60.

recommendations or decisions of the AI system must be either reversible or ignorable; and user data must be erased post-testing. Additionally, Article 60 mandates special protection for vulnerable groups, such as individuals with age-related vulnerabilities or physical or mental disabilities. Furthermore, the testing may be subject to unscheduled inspections by the authority to ensure compliance and safety.

2) Right to explanation of individual decision-making (Article 86)

Article 86 provides that any legal or natural persons are entitled to an explanation when a decision is made using an AI tool. This right to explanation applies when the decision is taken on the basis of the output of a high-risk AI system, and it has legal consequences or significantly affects the person in a manner they deem detrimental to their health, safety, or fundamental rights. The explanation provided must be clear and meaningful and provide the necessary information "on the role of the AI system in the decision-making procedure and the main elements of the decision taken." However, this provision applies solely to high-risk AI systems.

The specific benefits of this provision in relation to existing laws remain unclear, particularly because Article 22 of the General Data Protection Regulation (GDPR) already guarantees a right to an explanation for decisions made automatically. Nonetheless, this provision appears to extend the right to an explanation to include scenarios where a human decides "on the basis of the output" from an AI system, not just decisions "based solely on automated processing," as stipulated by Article 22 of the GDPR.

3) Articulation with the Digital Services Act (DSA)

Recital 118 of the AI Act highlights that some AI systems or models are embedded into Very Large Online Platforms or Very Large Online Search Engines designated by the EU Commission under the DSA regulation. As such, they are subject to the risk management framework provided for in the DSA. Therefore, the obligations of the AI Act—to assess and mitigate systemic risks—should be presumed to be fulfilled. In fact, the Commission has just requested information from a number of VLOPs and VLOSEs about the generative AI tools they use ((*see above section 5.1.1.D(2)*).

FIGURE 33. GPAI models embedded in VLOPs and VLOSEs

	Provisions applicable to GPAI models embedded in VLOPs and VLOSEs
Recital 118	● GPAI systems or models embedded into designated VLOPs or VLOSEs are subject to the risk management framework provided for in Regulation (EU) 2022/2065 (Digital Services Act). ● Consequently, the corresponding obligations of the AI Act should be presumed to be fulfilled unless significant systemic risks not covered by the DSA emerge and are identified in such models.

5.1.2.E Enforcement, sanctions, entry into force

The following chapters of the AI Act—specifically Chapter VII (Governance), Chapter XII (Penalties), and Chapter XIII (Final Provisions)—encompass provisions regarding the enforcement, sanctions, and entry into force of the Regulation.

1) Enforcement

Enforcement of the AI Act is the responsibility of both individual Member States and EU authorities.[1220] These authorities have specific responsibilities in the implementation of the provisions related to GPAI models. Additionally, it is important to highlight the significance of codes of practice in the enforcement of the Regulation and to mention the AI Pact initiative.

a) Competent authorities

i) National authorities[1221]

Each European Union Member State will appoint national authorities to supervise the enforcement of the AI Act within their respective jurisdictions. These authorities are to include at least one notifying authority[1222] and at least one market surveillance authority.[1223] Each Member State should also appoint one market surveillance authority to act as the single point of contact for the implementation of the Act.[1224] This authority will also represent the country on the European Artificial Intelligence Board. The AI Act leaves to each Member State the decision of whether to designate an existing independent authority to supervise the enforcement of the AI Act or establish a new agency for this purpose. In this context, data protection authorities emerge as strong candidates for the role due to their relevant expertise. However, Member States may also consider creating a new, independent, and autonomous agency.

While national authorities can act on their own initiative pursuant to the powers conferred upon them by domestic law, they can also be approached by any interested party. The AI Act grants any natural or legal person the right to lodge a complaint to the relevant market surveillance authority when they have reasons to consider that there has been an infringement of the provisions of the Act.[1225]

One of the noteworthy powers of market surveillance authorities is the ability to request access to the source code of a high-risk AI system.[1226] This access is granted when it is deemed necessary to evaluate compliance

---

[1220] Claudio Novelli et al., *A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities* (May 5, 2024), https://ssrn.com/abstract=4817755 or http://dx.doi.org/10.2139/ssrn.4817755*; See also* Kai Zenner, *The EU AI Act: responsibilities of the European Commission*, https://www.kaizenner.eu/post/ai-act-responsibilities-commission (last visited July 12, 2024)

[1221] AI Act, art. 70.

[1222] Notifying authorities are responsible for setting up and carrying out the necessary procedures for the assessment, designation, and notification of conformity assessment bodies and for their monitoring (AI Act, art. 30).

[1223] AI Act, art. 70(1).

[1224] *Id.* art. 70(2).

[1225] *Id.* art. 85.

[1226] *Id.* art. 74(13).

and when testing or auditing procedures, along with verifications based on the data and documentation provided by the provider, have been completed or found inadequate.

ii) EU Commission and AI Office

While for AI systems, the market surveillance system on the national level will apply, the provisions applicable to GPAI models are enforced at the EU level. According to Article 88, the European Commission holds exclusive authority to oversee and enforce the provisions applicable to general-purpose AI models. National market surveillance authorities can ask the Commission to exercise its powers to assist with the fulfillment of their own tasks, when it is necessary and proportionate. The Commission will also be responsible for adopting a number of implementing and delegated acts.[1227] It will, for instance, be responsible for specifying the criteria to be taken into account when designating GPAI models with systemic risk and adjusting various regulatory parameters. The Commission will also develop guidelines, for example, regarding prohibited practices.

The European Commission may delegate its enforcement authority to the newly established European Artificial Intelligence Office (AI Office), which was established within the Commission by a decision of on January 24, 2024.[1228] The AI Office is part of the administrative structure of the Directorate-General for Communication Networks, Content, and Technology. It is tasked with developing expertise and capabilities in artificial intelligence within the European Union and plays a central role in implementing EU legislation related to artificial intelligence.

The European AI Office is responsible for ensuring coordination of AI policy and fostering collaboration between EU institutions, bodies, agencies, experts, and stakeholders. It aims to establish a robust connection with the scientific community. It must play a supportive role to the Commission by aiding in decision-making and the adoption of delegated acts. And it must contribute to drafting guidelines and developing tools that facilitate these processes.

Moreover, within the context of the AI Act's implementation, the AI Office is given specific responsibilities:

- **Enforcement of provisions governing high-risk AI systems.** The AI Office may develop and recommend voluntary model contractual terms between providers of high-risk AI systems and third parties that supply tools, services, components, or processes that are used or integrated in high-risk AI systems (Article 28). The AI Office must also develop a questionnaire template to help draft fundamental rights impact assessments.

- **Enforcement of provisions governing AI systems based on GPAI models**. Among other tasks, the AI Office acts as a market surveillance authority in cases where an AI system is based on a general-purpose AI model and the model and system are provided by the same provider.[1229]

[1227] For a list *see* Zenner, *supra* note 1220.

[1228] Commission Decision Establishing the European AI Office (Jan. 24, 2024), EUROPEAN COMMISSION https://digital-strategy.ec.europa.eu/en/library/commission-decision-establishing-european-ai-office.

[1229] AI Act, art. 75.

- **Drafting of codes of practices**. The AI Office is tasked with encouraging and facilitating the drafting of codes of practice to facilitate the effective implementation of the provisions regarding the detection and labeling of synthetic content or the requirements related to GPAI models.

iii) The European Artificial Intelligence Board[1230]

To ensure uniform application of the AI Act, the European Artificial Intelligence Board, made up of one representative from each Member State, will meet to tackle extended tasks in advising and assisting the Commission and Member States. The Board will assume responsibility for various advisory roles, encompassing the issuance of opinions, recommendations, and advice, as well as contributing to the development of guidelines concerning the enactment of the AI Act. These responsibilities extend to offering insights on enforcement issues, technical specifications, and prevailing standards pertinent to the mandates outlined in the Act.

Additionally, the Board is tasked with furnishing advice to the Commission, the Member States, and their respective competent national authorities. The AI Act directs the Board to engage in collaboration, when deemed appropriate, with pertinent EU entities, expert panels, and networks that operate within the scope of relevant EU legislation, especially those involved in regulations concerning data, digital products, and services.

iv) Advisory forum[1231]

To provide additional technical expertise, the Commission will establish an advisory forum to advise and provide technical expertise to the AI Board and the Commission. This forum will represent a diverse group of stakeholders, including industry representatives, startups, small- and medium-sized enterprises (SMEs), civil society, and academia. The Commission will appoint advisory forum members from stakeholders with recognized expertise in artificial intelligence.

v) Scientific panel of independent experts[1232]

The Commission will also establish a scientific panel of independent experts who the Commission will select based on their up-to-date scientific or technical expertise in the field of artificial intelligence. This scientific panel will advise the AI Office, especially in the implementation and enforcement of the rules related to general-purpose AI models and systems. The scientific panel will provide expert advice on categorizing general-purpose AI models according to their systemic risk and contribute to the classification of various AI models and systems. The scientific panel will aid in developing tools and methodologies for assessing the capabilities of GPAI models and systems, including the creation and application of benchmarks. The panel will assist in formulating tools and templates to support these activities. And in

[1230] *Id.* art. 65-66.
[1231] *Id.* art. 67.
[1232] *Id.* art. 68-69.

parallel, the panel will extend its support to market surveillance authorities, responding to their requests for assistance. Finally, the panel will be tasked with alerting the AI Office of potential systemic risks for general-purpose AI models at the Union level.

b) Enforcement of provisions targeting GPAI models and systems

The rules governing GPAI models are enforced at the EU level. Each competent authority has a clearly defined mandate for enforcing the provisions related to GPAI models.

i) The EU Commission (or the AI Office acting as its delegate)

The European Commission holds exclusive authority to oversee and enforce the provisions applicable to general-purpose AI models, but it may delegate its authority to the AI Office.[1233] The powers conferred on the Commission (and the AI Office acting as its delegate) are as follows.

- **Power to request information**:[1234] The European Commission (or the AI Office acting as its delegate) can request documentation or additional information from providers of general-purpose AI models to assess compliance with the AI Act. Before making such requests, the AI Office may engage in dialogue with the provider. Moreover, the Commission has the authority to require information essential for the accomplishment of the scientific panel's objectives, upon request by the scientific panel. In such instances, requests must specify the legal basis, purpose, and required information, and set a deadline. Failure to provide requested information may result in fines.

- **Power to require access to GPAI models, including source code**:[1235] In the context of evaluation of GPAI models carried out by the AI Office, the Commission (or the AI Office acting as its delegate) can require access to a general-purpose AI model, including source code, through appropriate technical means, such as APIs. Before the request is made, the AI Office may engage in structured dialogue with the provider to gather more information on the internal testing of the model, internal safeguards for preventing systemic risks, and other internal procedures and measures the provider has taken to mitigate such risks.[1236] Providers of GPAI models or their representatives must grant requested access, and failure to do so may result in fines.

- **Power to appoint independent experts**:[1237] The European Commission (or the AI Office acting as its delegate) may appoint independent experts, including from the scientific panel, to conduct evaluations on its behalf.

[1233] *Id.* art. 88.
[1234] *Id.* art. 91.
[1235] *Id.* art. 92(3).
[1236] *Id.* art. 92(7).
[1237] *Id.* art. 92(2).

- **Power to request measures**:[1238] The European Commission (or the AI office acting as its delegate) is empowered to request providers of general-purpose AI models to take necessary measures to comply with their obligations. It is also empowered to implement mitigation measures in cases of serious systemic risks identified through evaluation and take actions such as restricting market availability or recalling the model. Before a measure is requested, the AI Office may engage in structured dialogue with the provider of the GPAI model. If the provider commits to implementing mitigation measures during this dialogue, the Commission may, by decision, make those commitments binding, thereby concluding any further action.

ii) The AI Office

Concerning the obligations of general-purpose AI *model* providers, the AI Office has the authority to monitor the effective implementation and adherence to the Act, including compliance with approved codes of practice.[1239] When the information obtained in response to a request by the Commission is insufficient, the AI Office, upon consultation with the Board, is empowered to conduct an evaluation of a general-purpose AI model to determine its provider's compliance.[1240] Additionally, the AI Office may perform evaluations to investigate systemic risks at the Union level associated with general-purpose AI models with systemic risk, especially following a qualified report from the scientific panel.[1241] Finally, the AI Office is tasked with creating a template for the detailed summary of the content used in training GPAI models.[1242]

Concerning AI *systems* based on general-purpose AI models, the AI Office acts as a market surveillance authority in cases where the general-purpose AI model and system come from the same provider.[1243] In other cases, national market surveillance authorities act as supervising authorities. They must cooperate with the AI Office to supervise general-purpose AI systems that can be used directly by deployers for at least one purpose that is classified as high-risk, when there is sufficient reason to believe the system is not compliant.[1244] In such a case, national authorities and the AI Office will carry out compliance evaluations and inform the Board and other market surveillance authorities accordingly. Moreover, national market surveillance authorities can request assistance from the AI Office when they are unable to conclude an investigation on a high-risk AI system because of their inability to gain access to certain information related to the general-purpose AI model. In such cases, the AI Office can take the necessary steps to make the information available.

iii) Scientific panel[1245]

[1238] *Id.* art. 93.
[1239] *Id.* art. 89(1).
[1240] *Id.* art. 92(1)(a).
[1241] *Id.* art. 92(1)(b).
[1242] *Id.* art. 53(1)(d).
[1243] *Id.* art. 75.
[1244] *Id.* art. 75(2).
[1245] *Id.* art. 90.

The scientific panel can issue a qualified alert to the AI Office if it suspects that a general-purpose AI model meets the requirements to be classified as a general-purpose AI model *with systemic risk* or poses a concrete identifiable risk at the Union level. Upon receiving such an alert, the Commission, through the AI Office and after informing the Board, can use its powers to assess the situation.

iv) Downstream providers[1246]

Downstream providers have the right to file a complaint for alleged violation of the AI Act by the provider of a general-purpose AI model. The complaint must be substantiated.

c) Codes of practice and harmonized standards

The development and enforcement of codes of practice are key elements in the implementation of the AI Act, especially for providers of GPAI models. In fact, many provisions of the AI Act refer to the implementation of codes of practice defined at Article 56 (*see Figure 34 below*). These codes will be developed with stakeholders from industry, the scientific community, civil society, and the EU Commission. Through an implementing act, the EU Commission may choose to approve a code of practice, granting it general validity within the European Union. Therefore, upon their development and approval, these codes of practices will enable providers of GPAI models to demonstrate their adherence to the AI Act, mirroring the approach adopted in the GDPR. If a code of practice cannot be finalized or is deemed inadequate by the AI Office when the AI Act becomes applicable, the Commission may establish common rules for the implementation of the relevant obligations.

Moreover, compliance with harmonized standards, which are generally expected to reflect the state of the art, should serve as a means for providers to demonstrate conformity with the requirements of the AI Act. "Harmonized standards" is a concept known from EU product safety legislation. A harmonized standard is a technical specification developed by a recognized European Standards Organization, such as the European Committee for Standardization (CEN), the European Committee for Electrotechnical Standardization (CENELEC), or European Telecommunications Standards Institute (ETSI).[1247] It is established following a request from the European Commission to one of these organizations. Once a harmonized standard has been approved, it can be used as a way of establishing a presumption of compliance with certain acts of EU law.[1248] The creation of AI-specific technical standards in collaboration with stakeholders should play a key role in providing technical solutions for providers to ensure adherence to the Act, as outlined by Recital 121.[1249] Compliance with these harmonized standards will provide AI providers with a legal presumption of conformity.

[1246] *Id.* art. 89(2).

[1247] *Harmonised Standards*, EUROPEAN COMMISSION https://single-market-economy.ec.europa.eu/single-market/european-standards/harmonised-standards_en#:~:text=A%20harmonised%20standard%20is%20a,to%20one%20of%20these%20organisations (last visited June 20, 2024).

[1248] The references of harmonized standards must be published in the Official Journal of the European Union (OJEU).

[1249] *Artificial Intelligence and Cybersecurity*, EUROPEAN PARLIAMENTARY RESEARCH SERVICE (Apr. 2024), https://www.europarl.europa.eu/RegData/etudes/ATAG/2024/762292/EPRS_ATA(2024)762292_EN.pdf.

The development of standardization is already underway. In May 2023, the European Commission directed CEN and CENELEC to formulate standards supporting the AI Act,[1250] setting an April 2025 deadline. Significant work has already been undertaken, particularly within the CEN-CENELEC JTC21 Special Advisory Group.[1251] Other standards-developing bodies, including the European Telecommunications Standards Institute (ETSI) and the International Organization for Standardization (ISO), are in the process of developing AI standards. In June 2023, the EU Agency for Cybersecurity (ENISA) released a multilayer security framework known as the Framework for AI Cybersecurity Practices (FAICP).[1252]

FIGURE 34. Codes of practice

	Codes of practice
Article 56(1) and (3) **Development of codes of practice**	• AI Office shall encourage and facilitate the drawing up of codes of practice. • AI Office may invite all providers of general-purpose AI models, as well as relevant national competent authorities, to participate in the drawing up of codes of practice. • Civil society organizations, industry, academia, and other relevant stakeholders, such as downstream providers and independent experts, may support the process.
Article 56(2) **Content of codes of practice**	Codes of practice cover at least the obligations of GPAI models and GPAI models with systemic risk, including: • the means to ensure that the technical information and the documentation for providers is kept up to date in light of market and technological developments; • the adequate level of detail for the summary about the content used for training; • the identification of the type and nature of the systemic risks at Union level, including their sources, where appropriate; and • the measures, procedures, and modalities for the assessment and management of systemic risks.
Article 56(7) **Adherence to codes of practice**	• AI Office may invite all providers of general-purpose AI models to adhere to the codes of practice.

[1250] Commission Implementing Decision of 22 May.2023 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence, C(2023)3215 – Standardisation request M/593, (May 22, 2023), https://ec.europa.eu/growth/tools-databases/enorm/mandate/593_en.

[1251] Josep Soler Garrido, et al., *Analysis of the preliminary AI standardisation work plan in support of the AI Act* Publications Office of the European Union, Luxembourg, 2023, https://publications.jrc.ec.europa.eu/repository/handle/JRC132833.

[1252] *Multilayer Framework for Good Cybersecurity Practices for AI*, Enisa, (June 7, 2023) https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai.

Article 56(6) **Implementation of codes of practice**	• AI Office and Board shall monitor compliance. • Commission may, by way of an implementing act, approve a code of practice and give it a general validity within the Union.
Article 53(4) and Article 55(2) **Codes of practice, harmonized standards and compliance**	• Providers of GPAI models and GPAI models with systemic risk **may rely on codes of practice** to demonstrate compliance with their obligations, until a harmonized standard is published. • Compliance with **European harmonized standards** grants providers the presumption of conformity to the extent that those standards cover those obligations. • Providers of general-purpose AI models who do not adhere to an approved code of practice or do not comply with a European harmonized standard shall demonstrate **alternative adequate means of compliance** for assessment by the Commission.

d) AI Pact

In anticipation of the implementation of the AI Act, the European Commission launched the AI Pact,[1253] initiating a call for voluntary commitments from companies both within and outside the EU. The pact aims to encourage the proactive adoption of measures outlined in the AI Act ahead of the legal deadline, emphasizing the responsible design, development, and use of AI technologies. The pact provides that companies must make concrete pledges detailing specific actions aligned with the AI Act's requirements.

The first call for interest in the AI Pact was launched in November 2023, receiving responses from over 550 organizations of various sizes, sectors, and countries. The AI Office has since initiated the development of the AI Pact, structured around two pillars: Pillar I serves as a gateway to engage the AI Pact network—those organizations that have expressed interest in the Pact. It encourages the exchange of best practices and provides practical information on the AI Act implementation process. Pillar II motivates AI system providers and deployers to prepare early and take proactive steps toward compliance with the requirements and obligations outlined in the legislation.

2) Sanctions[1254]

Member States hold the responsibility for setting the rules concerning penalties and other enforcement mechanisms, which may include warnings and non-monetary measures. However, in the case of noncompliance with specific provisions, the Act sets out maximum amounts based on the severity of the infraction.

[1253] AI Pact (last visited June 20, 2024), https://digital-strategy.ec.europa.eu/en/policies/ai-pact.
[1254] AI Act, art. 99.

- The Act calls for fines of up to 35 million Euros or 7% of a company's total worldwide annual revenues of the preceding financial year (whichever is higher) in case of noncompliance with a ban on prohibited AI practices,[1255]

- Fines should be up to 15 million Euros or 3% of the total worldwide annual revenues of the preceding financial year:

    - in case of noncompliance with the transparency obligations for chatbots, deepfakes, generative AI tools, emotion recognition systems, or biometric classification systems;[1256]

    - in cases where providers, authorized representatives, importers, distributors, deployers of high-risk AI systems, and notified bodies do not comply with their obligations;[1257] and

    - in cases where providers of GPAI models do not comply with the provisions of the Act or fail to comply with the requests made by the European Commission.[1258]

- Fines should be up to 7.5 million Euros or 1% of the total worldwide annual revenues of the preceding financial year in cases where a company supplies incorrect, incomplete, or misleading information to competent authorities and notified bodies.[1259]

The potential for significant fines and compliance costs under the AI Act could inadvertently benefit large, established technology companies, who can better afford to bear these costs than smaller companies. The AI Act provides that, for small- and medium-sized enterprises (SMEs), including startups,[1260] the threshold should be the lower of the two amounts mentioned in the provisions of Article 99.

3) Entry into force and application[1261]

The AI Act was published in the *Official Journal of the European Union* on July 12, 2024. On the 20th day following its publication -i.e. on August 1, 2024- the AI Act entered into force, as provided by Article 113. And 24 months after this date, on August 2, 2026, most parts of the Act will become fully enforceable. However, some deadlines are slightly shorter. The phased-in implementation is provided as followed by Article 113:

- Six months after entry into force (February 2, 2025): Prohibited AI practices will be effectively banned.

[1255] *Id.* art. 99(3).
[1256] *Id.* art. 99(4)(g).
[1257] *Id.* art. 99(4)(a to f).
[1258] *Id.* art. 101.
[1259] *Id.* art. 99(5).
[1260] *Id.* art. 99(6).
[1261] *Id.* art. 111 and 113.

- Nine months after entry into force (May 2, 2025): Codes of practice must be ready "in view of enabling providers to demonstrate compliance on time."[1262]
- 12 months after entry into force (August 2, 2025): The provisions concerning GPAI models, notifying authorities and notified bodies, governance, confidentiality, and penalties will go into effect, with the exception of the provisions relating to fines imposed on providers of general-purpose models (which means that no fines will be imposed for any violation of the requirements on GPAI models for another 12 months, creating an additional grace period).
- 24 months after entry into force (August 2, 2026): All other provisions of the AI Act will be enforceable, except the obligations for high-risk systems detailed in Annex I (*see above section 5.1.2.B(2))*.
- 36 months after entry into force (August 2, 2027): The obligations for high-risk systems detailed in Annex I (*see above section 5.1.2.B(2))* become applicable.

Article 111 provides exceptions to these principles for AI systems that are already on the market when the provisions of the AI Act become enforceable. Specifically, existing high-risk AI systems are exempted, which introduces a significant gap in AI safety.

- Providers of GPAI models that have been placed on the market before the rules on GPAI models are enforceable (August 2, 2025) shall comply with their obligations by 36 months after entry into force of the AI Act (August 2, 2027).[1263]

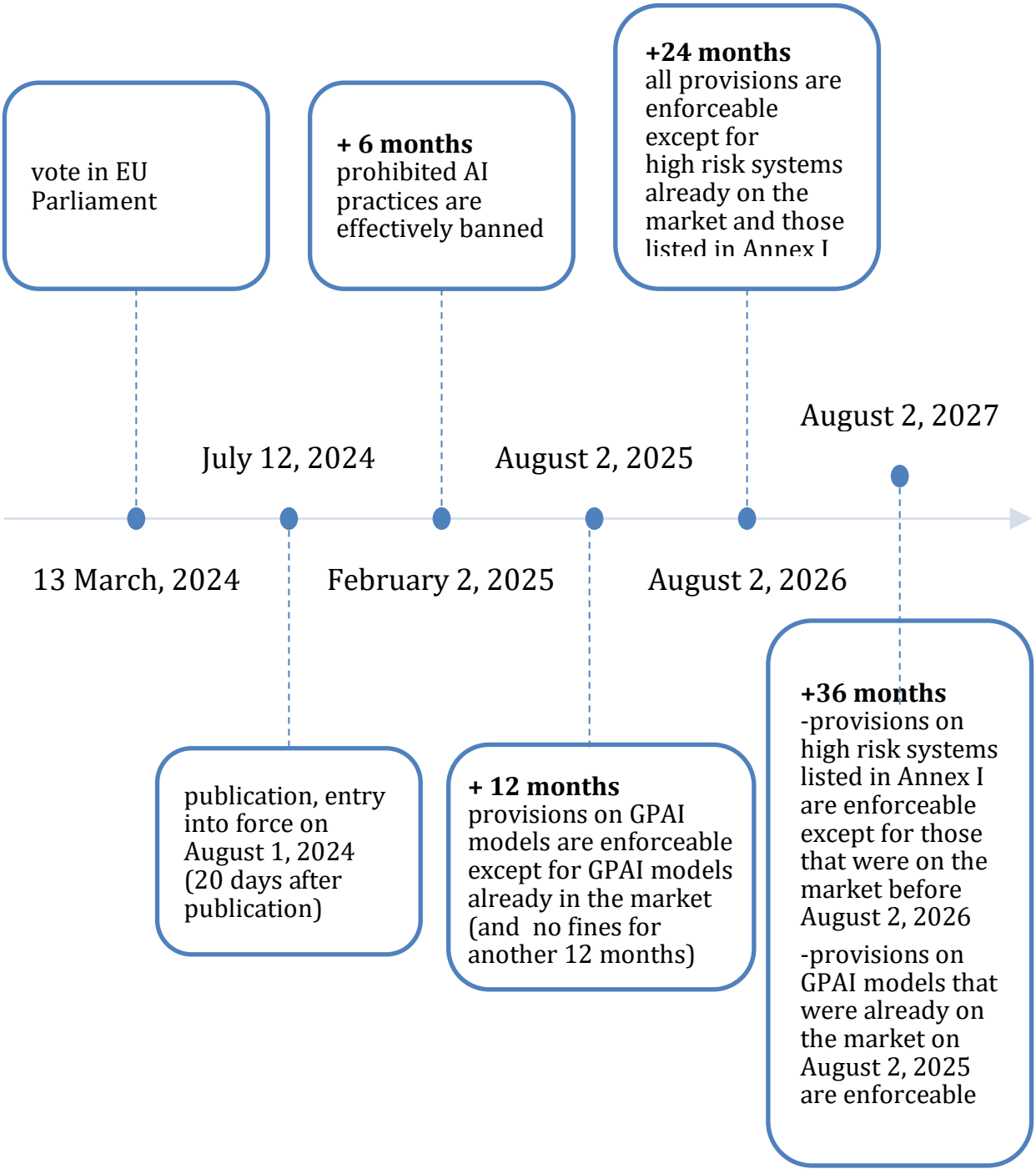
- Operators of high-risk AI systems that have been placed on the market or put into service *before* the AI Act becomes fully enforceable (August 2, 2026) will have to comply with the Act only if, from that date, their systems have been subjected to significant changes in their designs.[1264] Recital 177 provides that the concept of "significant change" should be "understood as equivalent in substance to the notion of substantial modification."

- AI systems that are components of large-scale IT systems that have been placed on the market or put into service *before* the date of 36 months after entry into force (August 2, 2027) shall be brought into compliance with the Act by December 31, 2030.[1265]

The timeline below provides an overview of the phasing-in of the AI Act.

FIGURE 35: Timeline of the entry into application of the AI Act

[1262] *Id.* Recital 179; art. 56(9).
[1263] *Id.* art. 111(3).
[1264] *Id.* art. 111(2).
[1265] *Id.* art. 111(1).



5.1.2.F. Conclusion on the AI Act

The AI Act is an ambitious initiative designed to create a comprehensive legislative framework for regulating AI, adopting a risk-based approach modeled after European product safety legislation. Several key aspects stand out. First, the AI Act combines two different approaches: regulating use cases and regulating the technology itself. Second, it attempts to address most of the risks associated with AI examined in Chapter 3. Third, enforcement will be particularly complex. Finally, the Act's extraterritorial impact is significant. It not only affects AI services within the EU but also, as the first comprehensive AI legislation, is poised to influence standards adopted globally.

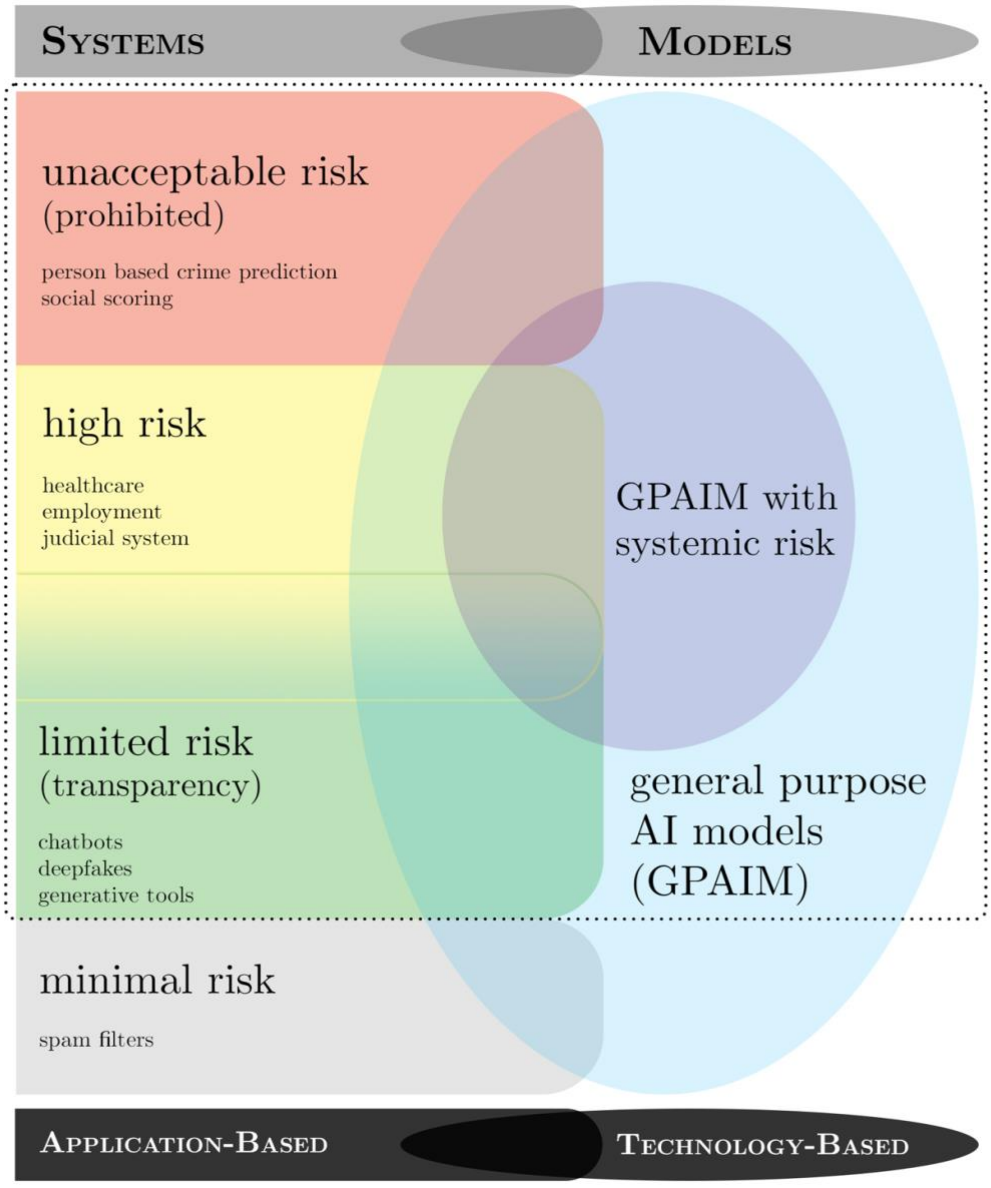
1) A dual approach

The AI Act exemplifies a dual logic in its regulatory approach. First, as originally proposed by the European Commission, it implements regulations by classifying AI systems into various risk categories according to their types and uses. Second, it introduces a supplementary layer of rules that concentrate on the technology itself, specifically targeting General-Purpose AI (GPAI) models. This additional regulatory layer targets *models*, rather than systems, but GPAI *systems* are also governed under the first layer of rules.

Within this framework, the AI Act's categories are not mutually exclusive but rather overlapping. For instance, AI applications that necessitate special transparency measures, such as chatbots, might also fall under the high-risk category if used in sectors listed in Annex III, like education or justice. In this case, the applicable rules are cumulative. Moreover, GPAI models can support AI systems that can be considered "high risk" or "transparency risk." Consider the forthcoming GPT-5, a GPAI model identified as having systemic risk and thus subject to corresponding regulations. Any chatbot developed using this model will also need to adhere to the transparency requirements specifically mandated for chatbots.

The figure below aims to depict both the dual logic and the overlapping scope of the regulations established by the AI Act.

FIGURE 36: Scope of the various AI Act layers



Source: Florence G'sell/ Ben Rosenthal

2) Mitigating AI risks

The AI Act is designed to protect fundamental rights and mitigate AI risks. The table below (Figure 37) attempts to present how the AI Act framework addresses the risks examined in Chapter 3.

Overall, the vast majority of risks and challenges discussed in Chapter 3 are addressed by the AI Act and, more broadly, by other EU legal frameworks, such as the GDPR and the DSA. However, there are some limitations. The provisions related to systemic risks are currently somewhat vague due to a lack of specifics —specifics which are expected to be provided in future delegated and implementing acts. The scope of the provisions relaxing the rules on open-source general-purpose AI models remains limited.

Furthermore, while the AI Act's reference to the Text and Data Mining exception in the New Copyright Directive confirms it is applicable to the training of AI models, it does not resolve the uncertainties about what regime is applicable to works created by generative AI or with the assistance of generative AI.

Additionally, although the AI Act addresses the crucial issue of the environmental impact of generative AI, its provisions remain relatively non-restrictive in that regard. It requires the disclosure of the "known or estimated energy consumption" of a general-purpose AI model but, if a company does not know the actual energy consumption, the Act allows the company to submit an estimate based solely on the computational resources used to train the model.

FIGURE 37. How the AI Act addresses identified risks

Possible risks and challenges of generative AI identified in Chapter 3	**AI Act provisions**
Technical vulnerabilities (section 3.1.1)	● **High-risk AI systems** must comply with various requirements (*section 5.1.2.B (2)*): ○ Risk management (Article 9 AI Act) ○ Data quality and governance (Article 10 AI Act) ○ Comprehensive technical documentation (Article 11, Annex IV AI Act) ○ Consistent recordkeeping (Article 12 and 20, AI Act) ○ Transparency and provision of information to deployers (Article 13, AI Act) ○ Guarantee of human oversight (Article 14, AI Act) ○ Ensuring system accuracy, robustness, and cybersecurity (Article 15, AI Act) ● **Providers of GPAI models** must draw up and keep up-to-date ○ a **technical documentation**, which must include a general description of the model and relevant information about the process used for its development, including the technical means (e.g., instructions of use, infrastructure, tools) required for the GPAI model to be integrated in AI systems; the design specifications of the model and training process; and information on the data used for training, testing, and validation (Article 53(1)(a) and Annex XI AI Act, s*ee section 5.1.2.C.1.*), and ○ make available **information and documentation to providers** who intend to integrate the GPAI

	model into their own AI system, in order to enable providers of AI systems to have a good understanding of the capabilities and limitations of the GPAI model and to comply with their own obligations (Article 53(1)(b) And Annex XII AI Act, *see section 5.1.2.C.1.*). ● **Providers of GPAI models with systemic risk** must (Article 55, AI Act, *see section 5.1.2.C.2.*): ○ perform model evaluation, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risk; ○ assess and mitigate possible systemic risk; ○ keep track of, document, and report relevant information about serious incidents and possible corrective measures to address them; ○ ensure an adequate level of cybersecurity protection; and ○ provide additional information in the technical documentation, such as detailed descriptions of the evaluation strategies; detailed descriptions of the measures put in place to conduct internal and/or external adversarial testing and model adaptations, including alignment and fine-tuning; and detailed descriptions of the system architecture (Annex XI AI Act, see *section 5.1.2.C.2.*).
Factually incorrect content (section 3.1.2.)	● **Providers** must: ○ design their chatbots to make sure natural persons know they are interacting with an AI (Article 50(1) AI Act, *section 5.1.2.B (3)*), and ○ ensure that generative AI outputs are marked in a machine-readable format and detectable as artificially generated or manipulated (Article 50(2)). ● **Deployers** must disclose that content was AI generated or manipulated (Article 50(4) AI Act, *section 5.1.2.B (3)*), especially in the case of AI systems that generate or manipulate text published to inform the public on matters of public interest (Article 50(4) AI Act, *section 5.1.2.B (3)).*

	• **Providers of GPAI models with systemic risk** must assess and mitigate possible systemic risk (Article 55, AI Act, *section 5.1.2.C.2.*).
Opacity (section 3.1.3)	• **Providers of GPAI models** must draw up and keep up-to-date technical documentation, which must include a general description of the model and relevant information about the process used for its development (Article 53(1) AI Act, *section 5.1.2.C.1.*).
Misuse and abuse (section 3.2.1)	**Providers of GPAI models with systemic risk** must: • perform model evaluation, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risk (Article 55, AI Act, *section 5.1.2.C.2.*), and • assess and mitigate possible systemic risk (Article 55, AI Act, *section 5.1.2.C.2.*).
Misinformation and Disinformation (section 3.2.2)	• **Very Large Online Platforms and Search Engines** must assess and mitigate systemic risks (Article 34 and 35 of the Digital Services Act, *section 5.1.1.D*).
Bias and discrimination (section 3.2.3)	• **High-risk AI systems** that continue to learn must be developed to eliminate or reduce the risk of possibly biased outputs influencing input for future operations ("feedback loops") and should be addressed with appropriate mitigation measures (Article 15(4) AI Act, *section 5.1.2.B (2)(c)*). • **Providers of GPAI models** must disclose information about the data used, including type and provenance of data and curation methodologies, the number of data points, their scope and main characteristics; how the data was obtained and selected, as well as all other measures to detect the unsuitability of data sources and methods to detect identifiable biases (Article 53(1) and Annex XI AI Act, *section 5.1.2.C.1.*). • **Providers of GPAI models with systemic risk** must assess and mitigate possible systemic risk (Article 55, AI Act, *section 5.1.2.C.2.*)
Influence, overreliance and dependence (section 3.2.4)	• **AI systems that exploit vulnerabilities of individuals** (Article 5(1)(b)), deploy subliminal techniques (Article 5(1)(a)), or infer emotions of a natural person in the areas of workplace and educational institutions ((Article 5(1)(f))) are prohibited (*section 5.1.2.B (1)*).

Nascent capabilities (section 3.2.5)	• **Providers of GPAI models with systemic risk** must assess and mitigate possible systemic risk (Article 55, AI Act, *section 5.1.2.C.2.*).
Open source models (section 3.2.6.A)	• **Providers of AI models** that are not with systemic risk and are released under a free and open license that allows for the access, usage, modification, and distribution of the model, and whose parameters are made publicly available are exempted from the obligations to draw up technical documentation and provide information to downstream developers
Highly Capable Models (Section 3.2.6.B)	• More stringent rules apply to "**general purpose AI models with systemic risk**," which are defined as GPAI models with "high-impact capabilities" (Article 3(65) AI Act, *section 5.1.2.C.2.*). • A GPAI model is presumed to have high-impact capabilities if it is trained using computing power exceeding 10^25 floating point operations (FLOPs) (Article 51(1) and 51(2) AI Act, *section 5.1.2.C.2.*). • A GPAI model may be designated as "with systemic risk" by the Commission on the basis of the following criteria: number of parameters, quality or size of the dataset, input and output modalities of the model, benchmarks and evaluations of capabilities of the model; degree of autonomy and scalability, tools it has access to, impact on the internal market due to its reach presumed if available to at least 10,000 registered business users), number of registered end users (Annex XIII AI Act, *section 5.1.2.C.2.*).

Privacy and data protection (section 3.3.1)	• **GDPR provisions** (*section 5.1.1.A.*).
	• **Providers of GPAI models** must:
	  ○ disclose information about the data used, including type and provenance of data and curation methodologies and how the data was obtained and selected (Article 53(1) and Annex XI AI Act, *section 5.1.2.C.1.*), and
	  ○ draw up and make publicly available a sufficiently detailed summary about the content used for training of the GPAI model, especially to facilitate parties with legitimate interests to exercise and enforce their rights (Article 53(1)(d), *section 5.1.2.C.1.*).

Copyrights (section 3.3.2)	• **Providers of general-purpose AI models** must put in place a policy to comply with EU copyright law and

	especially the Text and Data Mining exception (Article 53(1) AI Act, *section 5.1.2.C.1.*). ● **Text and Data Mining exception**: Web scraping is permitted only on condition that the use of protected materials "has not been expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online" (Article 4(3) of Directive (EU) 2019/790, *section 5.1.1.B.*). ● **Providers of GPAI models** must draw up and make publicly available a sufficiently detailed summary about the content used for training of the GPAI model, especially to facilitate copyright holders to exercise and enforce their rights (Article 53(1)(d), s*ection 5.1.2.C.1.*)
Environmental impact (section 3.4.3)	● **Providers of GPAI models** must disclose information on the known or estimated energy consumption of the model (Article 53(1), Annex XI AI Act) (*section 5.1.2.C.1.*).

3) Enforcement issues

The enforcement and monitoring of the AI Act are likely to be particularly challenging (*see section 5.1.2.E*). Effective implementation depends on the Commission adopting numerous delegated and implementing acts, and the AI Office adopting various implementation measures. Furthermore, the Commission and the AI Office must collaborate effectively with various bodies established at the Union level to support the AI Act's implementation. Collaboration with competent national authorities will also be essential. Ensuring these collaborations work efficiently will be challenging, particularly since multiple regulatory authorities within Member States may be involved. Overall, the proliferation of competent authorities and bodies required for the Act's implementation risks significantly complicating its application.

4) Extraterritorial scope

Last but not least, it is crucial to highlight the extraterritorial scope of the AI Act, which governs all generative AI services accessible and used within the European Union. Consequently, most global AI companies offering these services on an international scale will need to comply with these standards if they wish to operate within the European Union. While it is conceivable that services offered to European customers could differ from those provided elsewhere, the obligations stipulated by the AI Act render such differentiation both challenging and expensive. For instance, the AI Act's requirements for copyright compliance primarily affect the training of models. Considering the development costs, it is improbable that developers will create AI models exclusively for Europe. Similarly, implementing watermarking solely for European users would be difficult. Overall, the AI Act seems to establish standards intended for global application.

5.1.3. The liability directives

Almost 18 months after the publication of its proposal for an AI Act, the European Commission published two proposals for directives aimed at modernizing the civil liability principles applicable to artificial intelligence.[1266] One of the two published September 28, 2022, the Proposal for a Product Liability Directive (PLD),[1267] updates the 1985 Product Liability Directive[1268] to explicitly include software and AI systems within its purview. The second proposal was a new AI Liability Directive (AILD),[1269] designed to ensure that "victims of damage caused by AI obtain equivalent protection to victims of damage caused by products in general," as stated by the proposed draft.

As described in the explanatory memorandum of the AILD proposal,[1270] liability ranks among the top three barriers to the use of AI by European companies. It is cited (by 43%) as the most relevant external obstacle for companies that are planning to—but have yet to—adopt artificial intelligence. The Commission's aim is, therefore, to provide legal certainty while ensuring that victims can be effectively compensated for their losses.

The proposed directives and the AI Act serve as complementary measures.[1271] The AI Act contains precise and comprehensive regulations, governing the development and deployment of AI technologies; the directives specifically address the rights of individuals adversely affected, providing provisions for their protection. As they did with the AI Act, the drafters of these directives sought to address the complexity and diversity of the supply chain, which encompasses a wide array of players, including developers, deployers, and users. Notably, they have tried to delineate responsibilities among the original developers, the deployers who fine-tune the models for specific tasks, and the front-end operators who use the systems in their daily activities. The texts of both directives offer a somewhat imperfect solution to these challenges.

[1266] Philipp Hacker, *The European AI liability directives – Critique of a half-hearted approach and lessons for the future*, Computer Law & Security Review, 51 Computer Law & Security Review, 1 (2023), https://doi.org/10.1016/j.clsr.2023.105871; *see also* Philipp Hacker, *The European AI Liability Directives—Critique of a Half-Hearted Approach and Lessons for the Future*, Oxford Business Law Blog (March 15, 2023), https://blogs.law.ox.ac.uk/oblb/blog-post/2023/03/european-ai-liability-directives-critique-half-hearted-approach-and-lessons.

[1267] Proposal for a directive on liability for defective products, COM/2022/495 final (**New Product Liability Directive**), (Sept. 28, 2022), art. 4(1) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0495; *see* European Parliamentary Research Service, *New Product Liability Directive* (Dec. 2023), https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739341/EPRS_BRI(2023)739341_EN.pdf.

[1268] Council Directive 85/374/EEC of July 25, 1985 (**Product Liability Directive**), O.J. (L 210/29), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31985L0374.

[1269] Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence, (**AI Liability Directive**), COM (2022) 496 final (Sept. 28, 2022), https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52022PC0496.

[1270] *Id.* at 1.

[1271] Philipp Hacker, *The European AI liability directives – Critique of a half-hearted approach and lessons for the future*, *see supra* note 1266.

This section will first examine the revised Product Liability Directive, adopted on March 12, 2024, and then consider the proposed AI Liability Directive, which is still pending adoption.

5.1.3.A. The revision of the Product Liability Directive

Following the release of Commission's Proposal on September 22, 2022, the European Council[1272] and the European Parliament[1273] each presented their views on the proposed Product Liability Directive. On December 14, 2023, the European Parliament and Council reached a provisional agreement on the draft,[1274] which was approved by the Committee of the Permanent Representatives of the Governments of the Member States to the European Union (COREPER) on January 24, 2024.[1275] The new Product Liability Directive was approved by the EU Parliament on March 12, 2024.[1276]

This section will briefly outline the original product liability framework set up by the 1985 Directive, before turning to the revised Directive, which now includes software within its scope. The following comments are based on the final text of the new Product Liability Directive adopted by the EU Parliament in March.[1277]

1) The previous Product Liability framework

The harmonization of product liability laws within the EU was achieved in 1985 through the 85/374/EEC Product Liability Directive,[1278] which introduced a no-fault liability framework ("strict liability"). Under this previous framework, the claimant needs to establish the occurrence of harm, identify a defect in the product, and show that this defect directly caused the harm. The manufacturer is then required to compensate the claimant; however, the claimant could still bring an action for negligence based on national law.

a) Is an AI system or model a product?

[1272] Council of the European Union, Proposal for a directive on liability for defective products - Mandate for negotiations with the European Parliament 2022/0302(COD) (June 14, 2023).
[1273] European Parliament, *Report on the proposal for a directive of the European Parliament and of the Council on liability for defective products* (Oct. 12, 2023), https://www.europarl.europa.eu/doceo/document/A-9-2023-0291_EN.html#_section6.
[1274] European Parliament, *Deal to better protect consumers from damages caused by defective products* (Dec. 14. 2023), https://www.europarl.europa.eu/news/en/press-room/20231205IPR15690/deal-to-better-protect-consumers-from-damages-caused-by-defective-products.
[1275] Council of the European Union, Proposal for a directive on liability for defective products - Confirmation of the final compromise text with a view to agreement, January 18, 2024 (2022/0302(COD)), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0132_EN.html#title2.
[1276] European Parliament legislative resolution P9_TA(2024)0132 of Mar. 12, 2024 on the proposal for a directive on liability for defective products (COM(2022)0495 – C9-0322/2022 – 2022/0302(COD)), (**New Product Liability Directive**), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0132_EN.html.
[1277] *Id.*
[1278] Directive 85/374/EEC, *supra* note 1268.

Until now, there has been no legal precedent where compensation has been awarded for damages resulting from a defect in an AI-based product. Although the Product Liability Directive seemed like it *could* cover such cases, there was uncertainty about whether software falls under the definition of "products" as described in the Directive. The Directive, specifically Article 2, was traditionally interpreted as defining a "product" as a tangible item. But the core of the discussion centered on the extent to which the directive's definition of a "product" *ought* to be construed, particularly given that software is not typically provided in tangible form. Certainly, these provisions could encompass software when it is delivered on a physical medium (such as a car that includes AI-enabled features) or embedded within hardware. The status of stand-alone software–for instance, downloadable applications, or software accessible through a web interface—remained significantly more ambiguous.

b) The defectiveness

Under the Product Liability Directive framework, a plaintiff has the challenging task of proving that a product is defective and that this defect directly caused the harm they suffered. A product is deemed defective if it does not provide a level of safety that people are entitled to expect, considering all relevant factors. This could relate to how the product is marketed or the expected uses of the product at the time it was released.

In determining if a product is defective, courts assess whether the general public would reasonably perceive the product as unsafe, instead of basing their judgment on the personal viewpoint of the individual who experienced harm. Assessing the expected level of security for generative AI tools poses a significant challenge in this context. Is it justifiable to assume that a model does not meet user safety expectations when it generates inappropriate content? As mentioned earlier, providers of these tools typically warn users about the possibility of producing inaccurate content, which inevitably influences user expectations.

c) Defenses

In terms of the defenses available to producers, the Product Liability Directive exempts them from liability if they can prove the product was free from defects at the time of release. This principle is based on the understanding that a manufacturer's obligation is to guarantee the safety of the product until it is sold and does not cover subsequent problems arising from third-party interventions. This concept is logical and straightforward for physical, mass-produced products that typically do not undergo changes after their sale. However, the situation is significantly more complicated for digital products, particularly those that integrate both hardware and software, and especially those utilizing artificial intelligence, such as machine-learning algorithms. These products can evolve and adapt over time as they process new data, which may originate from external sources. Although traditional manufactured goods might be minimally affected by external factors, the task of identifying defects in AI-driven products is notably more difficult due to their dynamic and evolving nature.

Another significant circumstance that exempts AI model producers from liability occurs when they can prove that, at the time the product was introduced to the market, the prevailing level of scientific and technical knowledge did not allow for the identification of the defect. This is referred to as the defense of development risk.

2) The revision of the Product Liability Directive

The revised Product Liability Directive (PLD) explicitly encompasses software, including AI applications, within its scope. It maintains the traditional definition of a defect while extending it to software. The list of recoverable damages has been slightly expanded. Finally, the new Directive introduces new evidence rules.

a) Inclusion of software in the list of potentially defective products

The proposed revision of the Product Liability Directive broadens its applicability to clearly include software, whether or not it is integrated into a tangible product. All AI systems and AI-enabled goods are covered. However, the source code of software is not to be considered as a "product" within the meaning of the Directive, as it is pure information.[1279]

The scope of the revised Product Liability Directive does not include "free and open-source software that is developed or supplied outside the course of a commercial activity."[1280] "Free and open-source software" covers cases where "the source code is openly shared and users can freely access, use, modify and redistribute the software or modified versions thereof."[1281] Such software "is subject to licenses that allow anyone the freedom to run, copy, distribute, study, change and improve the software."[1282] Indeed, in such cases, the software cannot be considered as placed on the market.

However, this exemption does not apply "where software is supplied in exchange for a price or personal data are used other than exclusively for improving the security, compatibility or interoperability of the software, and is therefore supplied in the course of a commercial activity."[1283] In addition, the exemption does not apply if free and open-source software is subsequently integrated by a manufacturer as a component into a product in the course of a commercial activity. In such cases, the manufacturer could be liable for damage caused by the defectiveness of such software.

FIGURE 38.

Previous Product Liability Directive (1985)	**New Product Liability Directive (2024)**

[1279] New Product Liability Directive, Recital 13.
[1280] *Id.* art. 2(2).
[1281] *Id.* Recital 14.
[1282] *Id.*
[1283] *Id.*

Definition of product (Article 2)	Definition of product (Article 4(1))
• all movables, even if incorporated into another movable or into an immovable (Article 2)	• all movables, "even if integrated into or inter-connected with another movable or into an immovable" (Article 4(1)) • includes **software**, such as "operating systems, firmware, computer programs, applications or AI systems, (...) irrespective of the mode of its supply or usage, and therefore irrespective of whether the software is stored on a device or accessed through a communication network or cloud technologies, or supplied through a software as a service model" (Recital 13) • **Information** is not "to be considered as a product, and product liability rules should therefore not apply to the content of digital files, such as media files or eBooks or the **source code of software**" (Recital 13)
	Exemption for open-source software Article (2(2)) • the PLD does not apply to "free and open-source software that is developed or supplied outside the course of a commercial activity" (Article 2(2)) i.e. software (including its source code and modified versions) that is openly shared and freely accessible, usable, modifiable, and redistributable (Recital 14). • except where: • software is supplied in exchange for a price, • personal data is used other than exclusively for improving the security, compatibility, or interoperability of the software, and is, therefore, supplied in the course of a commercial activity.

b) Extension of the number of potential defendants to include AI systems providers

Under the revised PLD, developers or producers of software are considered manufacturers. Recital 13 expressly states that AI system providers, as defined by the AI Act, [1284] should be treated as "manufacturers."

[1284] AI Act, art. 3(2): 'provider' means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and

FIGURE 39. Liable operators under the Product Liability Directive

Previous Product Liability Directive (1985)	New Product Liability Directive (2024)
Liable operators (Article 3) • **Producer** is defined as "the manufacturer of a finished product, the producer of any raw material or the manufacturer of a component part and any person who, by putting his name, trademark or other distinguishing feature on the product, presents himself as its producer" (Article 3(1)) • EU importer[1285] (Article 3(2)). • Supplier who does not disclose, within a reasonable time, the identity of the person who supplied the product to them (Article 3(3)).	**Liable operators (Article 8)** ● **Manufacturer that:** - develops, manufactures, or produces a product or component, - put their name, trademark, or other distinguishing feature on the product ● Any person that "**substantially modifies** a product outside the manufacturer's control and thereafter makes it available" ● If the manufacturer is established outside the EU: - **importer** - **authorized representative** of manufacturer - **fulfillment service provider** ● **Distributor** (if no manufacturer can be identified), including online platform (such as a marketplace) when an average consumer could believe that the platform provides the product or that the product is provided by a trader acting under the platform's authority or control (as outlined by Article 6(3) DSA)

Third parties who make "substantial modifications" to a product after its market introduction bear liability, provided such modifications occur beyond the original manufacturer's control. The determination of whether a modification is substantial relies on criteria established in product safety legislation. In cases where specific criteria for a given product are not available, substantial modifications are described as "modifications that change the original intended functions of the product or affect its compliance with applicable safety requirements or change its risk profile."[1286]

places them on the market or puts the system into service under its own name or trademark, whether for payment or free of charge.

[1285] An importer is a person established in the EU that places a device from a third country on the EU market.

[1286] New Product Liability Directive, Recital 39.

Regarding AI products, the definition of the Directive should be put in light of the definition of "substantial modification" in the AI Act. According to Article 3(23) of the AI Act, a "substantial modification" refers to any change that was "not foreseen or planned in the initial conformity assessment" of the AI system. Recital 128 of the AI Act adds that the normal evolution that occurs as part of a machine-learning model's expected development does not count as substantial modifications. In this context, downstream users that employ general-purpose AI models and customize them for particular use cases may be liable. Nonetheless, deployers and users who only make minor adjustments to an existing AI model or system should not take on the responsibilities, liabilities, and compliance obligations that belong to the providers of these models. Of course, the issue centers on identifying the threshold at which a model modification is considered substantial.

c) Adapting the notion of defect to the particularities of software and AI models

Defectiveness is defined by the revised Product Liability Directive as a situation where "the product does not provide the safety that a person is entitled to expect or that is required under Union or national law."[1287] The required level of safety as mandated by the Union or a Member State's law is based on legally established safety standards. For AI systems, the provisions of the AI Act, particularly regarding high-risk AI systems or general-purpose AI systems, will be considered.

Another pivotal question is identifying the level of safety that a person should reasonably expect from an AI system. The new PLD provides detailed guidance on the factors to be considered in such an assessment.[1288]

FIGURE 40. Defectiveness under the Product Liability Directive

Previous Product Liability Directive (1985)	**New Product Liability Directive (2024)**
Defect (Article 6) The product does not provide the level of safety that a person is entitled to expect, taking all circumstances into account, including: • the presentation of the product (Article 6(1)(a)). • the use to which it could reasonably be	**Defect (Article 7)** 1. A product shall be considered defective when it does not provide the safety that a person is entitled to expect or that is required under Union or national law. 2. In assessing the defectiveness of a product, all circumstances shall be taken into account, including:

[1287] *Id.* art. 7(1).

[1288] Recital 22 states that the safety that the public at large is entitled to expect "should be assessed by taking into account, inter alia, the intended purpose, reasonably foreseeable use, the presentation**,** the objective characteristics and the properties of the product in question, including its expected lifespan**,** as well as the specific requirements of the group of users for whom the product is intended."

expected that the product would be put (Article 6(1)(b)). • the time when the product was put into circulation (Article 6(1)(c)).	*(a) the presentation and the characteristics of the product, including its labeling, design, technical features, composition and packaging, and the instructions for its assembly, installation, use, and maintenance;* *(b) the reasonably foreseeable use of the product;* *(c)* ***the effect on the product of its ability to continue to learn or acquire new features*** *after it is placed on the market or put into service;* *(d) the reasonably foreseeable effect on the product of other products that can be expected to be used together with the product, including by means of interconnection;* *(e) the moment in time when the product was placed on the market or put into service or, where the manufacturer retains control over the product after that moment, the moment in time when the product left the control of the manufacturer;* *(f) relevant product safety requirements, including safety-relevant cybersecurity requirements;* *(g) any recall of the product or any other relevant intervention by a competent authority or by an economic operator referred to in Article 7 relating to product safety);* *(h) the specific needs of group of users for whose use the product is intended;* *(g) in the case of a product whose very purpose is to prevent damage, any failure of the product to fulfill that purpose.*

Interestingly, the revised PLD provides in Recital 30 that "the assessment of defectiveness should involve an objective analysis of the safety that *the public at large* is entitled to expect**,** and not refer to the safety that any particular person is entitled to expect". However, does the public have sufficient awareness of the recognized risk of hallucinations inherent in models like ChatGPT, especially when it has been explicitly stated that these models can make errors? Recital 31 specifies that "warnings or other information provided with a product cannot be considered sufficient to make an otherwise defective product safe, since defectiveness should be determined by reference to the safety that the public at large is entitled to expect." Thus, one could conclude that merely warning users that a model might hallucinate, produce inaccurate information, or behave unpredictably does not suffice to exempt that product from being deemed defective. On the other hand, would it be reasonable to consider that the public expects that such models never produce inaccurate or illegal content? For the moment, there do not seem to be any clear, definitive answers to these questions.

Moreover, it is worth mentioning that "the reasonably foreseeable use of the product"[1289] should be considered in the assessment. Does this suggest that the providers of generative AI chatbots should foresee

[1289] New Product Liability Directive, art. 7(2)(b).

that students, journalists, or lawyers will use their tool and rely on the information it generates? While the answer to that question is uncertain, it is evident that providers of AI systems must foresee potential misuse. Recital 31 provides that the determination "of reasonably foreseeable use should also encompass misuse that is not unreasonable under the circumstances, such as the foreseeable behaviour of a user of machinery resulting from a lack of concentration or the foreseeable behaviour of certain user groups such as children." Consequently, providers should expect that schoolchildren might use their tools for doing their homework or that teenagers might seek to create deepfakes.

The revised Product Liability Directive also mentions "the effect on the product of its ability to continue to learn or acquire new features after it is placed on the market or put into service."[1290] An AI system could be deemed defective due to knowledge acquired after its release. Recital 32 states that "a manufacturer that designs a product with the ability to develop unexpected behaviour should remain liable for behaviour that causes harm."

Once again, it is challenging to ascertain the full impact of such a provision at this stage. The text appears to hold developers of advanced AI models liable for the most unpredictable effects of their applications. In a context where it is inherently impossible to fully anticipate the behavior of machine-learning tools, this responsibility could be substantial.

d) Possible claimants and repairable damage

While the proposal substantially raises the number of potential defendants, it is relatively narrow in defining who can sue and what kind of compensation they can pursue. Only natural persons are entitled to compensation.[1291] This implies that legal entities, beginning with businesses, are not entitled to initiate legal proceedings based on product liability against a provider of an AI system.

The revised PLD extends the scope of recoverable damages to cover death, physical injury, or property damage (unless the property is used for professional purposes), as well as losses resulting from "medically recognized damage to psychological health"[1292] or "destruction or corruption of data that are not used for professional purposes."[1293]

FIGURE 41 Damages compensated under the Product Liability Directive

Previous Product Liability Directive (1985)	New Product Liability Directive (2024)
Damage (Article 9) • death, physical injury, or property damage (unless the property is used	**Damage (Article 6)** • death or personal injury • "medically recognized damage to psychological health"

[1290] *Id.* art. 7(2)(c).
[1291] *Id.* art. 5.
[1292] *Id.* art. 6(1)(a).
[1293] *Id.* art. 6(1)(c).

exclusively for professional purposes) (Article 9(a)) • minimum claims threshold (500 euros) (Article 9(b).	• "destruction or corruption of data that is not used for professional purposes" • damage to or destruction of any property (except the product itself, property used exclusively for professional purpose and a product damaged by a defective component) • no minimum claims threshold

The directive establishes a framework that solely advantages natural persons, specifically for personal and nonprofessional harm. Notably, it encompasses psychological damage, which holds significant implications for the realm of generative AI. Individuals depicted in deepfakes or targeted by incorrect information produced by AI could experience this form of injury. However, the directive requires that damage to psychological health must affect the general state of health of the victim and may necessitate therapy or medical treatment.[1294]

Furthermore, the question arises whether compensation for the destruction or corruption of data might apply in situations where an AI model regurgitates its training data. Recital 20 addresses scenarios involving the deletion of digital files, for example, and specifies that "destruction or corruption of data is distinct from data leaks or breaches of data protection rules."[1295] This indicates that these latter issues fall within the scope of the GDPR. Consequently, the regurgitation of personal data is governed by the GDPR, not the Directive. Instead, the revised PLD covers scenarios where a malfunction in an AI system results in data loss, rendering the data inaccessible. The data in question does not need to be personal data but needs to be used for noncommercial purposes. Overall, it is challenging to ascertain how and when using a generative AI tool might lead to the corruption or loss of an individual's data. At first glance, such hypotheses appear limited.

e) Adjusting rules of evidence and burden of proof for digital products and software

A plaintiff must bring evidence to establish the defect, the harm they have suffered, and the connection between the two. However, in the context of software and AI models, proving a violation of safety standards or a distinct product malfunction may necessitate an examination of intricate technical details. Hence, the revised PLD states that plaintiffs who present "facts and evidence sufficient to support the plausibility of the claim" are entitled to obtain "relevant evidence" from the defendant.[1296]

If the defendant does not comply, the defectiveness of the product is presumed.[1297] This new rule needs to be implemented into the national laws of Member States. Its significance should not be undervalued in legal systems that lack mechanisms for discovery or disclosure.

FIGURE 42 Evidence rules under the Product Liability Directive

[1294] *Id.* Recital (21).
[1295] *Id.* Recital 20.
[1296] *Id.* art. 9(1).
[1297] *Id.* art. 10(2)(a).

Previous Product Liability Directive (1985)	New Product Liability Directive (2024)
Disclosure of evidence *None*	**Disclosure of evidence** **Article 9** • Claimants who are able to present facts and evidence sufficient to support the plausibility of the claim will be entitled to the disclosure of "relevant evidence" that the defendant has at its disposal (Article 9(1)). • If the defendant fails to disclose relevant evidence, the defectiveness of the product is presumed (Article 10(2)(a)).

Another significant change is related to presumptions. National courts should presume a product's defectiveness or the causal link between the defect and the damage, or both, in various circumstances listed in the table below. In particular, courts can presume these elements when proving them is excessively difficult for the claimant due to technical or scientific complexity. Given that manufacturers possess expert knowledge and more information than the injured person, a fair risk apportionment requires that claimants demonstrate only the likelihood of defectiveness or causation. Recital 48 provides that national courts should determine technical or scientific complexity on a case-by-case basis, considering factors such as the product'' innovative nature, the technology involved, the complexity of the information and data to be analyzed, and the intricacy of establishing the causal link. Claimants must argue "excessive difficulties," but proof of such difficulties is not required. For instance, "in a claim concerning an AI system, the claimant should, for the court to decide that excessive difficulties exist, neither be required to explain the AI system's specific characteristics nor how those characteristics make it harder to establish the causal link."[1298]

FIGURE 43 Presumptions under the Product Liability Directive

Previous Product Liability Directive (1985)	New Product Liability Directive (2024)
No presumption	**Presumption of defectiveness (Article 10(2))** Rebuttable presumption of defectiveness if: ● the defendant does not disclose relevant evidence when applicable (Article 9(1), see above), ● the claimant demonstrates that the product is noncompliant with mandatory product safety requirements targeting the risk of damage suffered by the claimant, or ● the claimant demonstrates that the damage was caused by an obvious

[1298] *Id.* Recital 48.

	malfunction of the product during reasonably foreseeable use or under ordinary circumstances.
	Presumption of causal link (Article 10(3)) Rebuttable presumption of causal link if: • defectiveness has been established (including on the basis of the presumption provided by Article 10(2), and • the damage caused is of a kind typically consistent with the defect in question.
	Presumption of defectiveness and/or causal link (Article 10(4)) Rebuttable presumption of defectiveness or causal link if: • claimant "faces excessive difficulties, in particular due to technical or scientific complexity" (Article 10(4)(a)) to prove the defectiveness or the causal link, or • claimant demonstrates that it is likely that the product is defective or there is a causal link between the defectiveness and the damage, or both.

f) Defenses and exemptions

The revised Product Liability Directive (PLD) modifies the range of defenses and exemptions available to defendants (see table below). The Directive maintains the current exemptions, such as the development risk exemption, while also incorporating new clarifications and exemptions specifically designed for AI technologies. It is important to note that defendants are liable only for factors under their control. As a result, they may be exempt from liability if they demonstrate that a product defect was absent at the time of market release or that the defect emerged after release.

However, the revised PLD restricts this exemption in situations where the defendant maintained a degree of control. For example, if a product's defectiveness stems from the absence of necessary software updates or upgrades to ensure safety, the manufacturer will bear liability. If the defectiveness results from a substantial modification of the software that took place under the manufacturer's control, the manufacturer will also be held liable. However, if the substantial modification is made outside of the original manufacturer's control, the person who made the modification is liable.

Implementing these principles for machine-learning algorithms that continue to learn after deployment will be challenging. Recital 50 indicates that manufacturers should be accountable for defects that emerge after the product has been made available on the market if these defects arise "as a result of software or related

services within their control, be it in the form of upgrades or updates or machine-learning algorithms." Software or related services are deemed under the manufacturer's control if they are provided directly by the manufacturer, or if the manufacturer approves, authorizes, or consents to their provision by a third party.

FIGURE 44 Liability exemption under the Product Liability Directive

Previous Product Liability Directive (1985)	**New Product Liability Directive (2024)**
Exemption from liability (Article 7) • Defendant did not put the product into circulation (Article 7(a)). • It is probable that the defect that caused the damage did not exist at the time when the defendant put the product into circulation (Article 7(b)). • Defendant did not manufacture the product for sale or any form of distribution for economic purposes nor did defendant manufacture or distribute the product in the course of their business (Article 7(c)). • The defect is due to compliance with mandatory regulations issued by public authorities (Article 7(d)). • "Development risks" or "state-of-the-art" defense (optional for Member States): The state of scientific and technical knowledge at the time the product was put into circulation was not such as to enable the discovery of the defect (Article 7(e)). • In the case of a manufacturer of a component, the defect is attributable to the design of the product in which the component has been fitted or to the instructions given by the manufacturer of the product (Article 7(f)).	**Exemption from liability (Article 11 (1))** • Defendant did not put the product in the market or into service. • Probable that defectiveness did not exist when the product was released, or probable that it came into being after release. • The defectiveness is due to compliance with legal requirements. • "Development risks" or "state-of-the-art" defense" (not optional): the defectiveness could not be discovered given the objective state of scientific and technical knowledge at the time the product was released, or during the time the product was within the manufacturer's control. • The defectiveness of the component is attributable to the design of the product in which it has been integrated or to the instructions given by the manufacturer of that product. • The defectiveness is related to a part of the product not affected by the modification realized by the defendant (in case the product was substantially modified)
	Exception to the exemption from liability (Article 11(2)) Defendant is still liable if the defectiveness: • is due to a related service, a software (including software update/upgrades), a lack of software updates/upgrades necessary to maintain safety, or a substantial modification, • provided that these were within the manufacturer's control.

3) Conclusion on the revised Product Liability Directive

Overall, the new Product Liability Directive significantly expands the liability of AI providers. However, the modest expansion of compensable damages and the limitation on who can sue significantly mitigate the increased liability of AI developers resulting from the definition of defectiveness. Since only natural persons can claim compensation for the damages explicitly listed in the text, the liability risk for AI product manufacturers appears to remain manageable.

Moreover, concern arises from the possibility that Member States and national courts may have varying opinions on what constitutes an acceptable level of safety, potentially resulting in legal uncertainty for providers of generative AI systems.

In any case, the directive's expansive notion of defect and the intent to hold manufacturers accountable for risks arising from the product's autonomous evolution may influence judicial interpretation of the general principle of fault-based liability under ordinary law. In this regard, the directive marks a significant step toward broader liability, both temporally and geographically, for AI developers.

5.1.3.B. The new proposal for an AI Liability Directive (AILD)

On the same day that it introduced its proposal to revise the Product Liability Directive (PLD), the European Commission unveiled a proposal for a new Artificial Intelligence Liability Directive (AILD). This proposal does not create a new liability framework but essentially mandates that Member States enact legislation substantially easing the burden of proof for individuals seeking to file civil liability lawsuits because they were harmed by AI-related products or services. It also provides for a "presumption of causality" that attributes liability to the provider or deployer/user of the AI system in certain circumstances. As is the case with the revised PLD, this new AI Liability Directive is intended to be implemented in tandem with the EU AI Act. However, as for now, only the PLD has advanced through the legislative process. The following presentation is based on the proposal released by the EU Commission on September 28, 2022.[1299]

The Commission started from the observation that claims related to harm caused by AI systems present significant challenges due to the complexity and lack of transparency of such systems. Claimants frequently face difficulties in establishing causality, largely because they lack access to the inner functioning of the AI system, a phenomenon often referred to as a "black box."[1300] Consequently, claimants may find it impossible to establish that a particular flaw or malfunction within the AI system directly caused their

[1299] Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence, (**AI Liability Directive - AILD**), COM (2022) 496 final (Sept. 28, 2022), https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52022PC0496.

[1300] Explanatory Memorandum, Proposal for a directive adapting the rules on non-contractual civil liability to the field of artificial intelligence (AILD), *see supra* note 1299, at 1

harm.[1301] The Commission has observed that individuals harmed by AI may face substantial initial costs and endure considerably lengthier legal processes compared to cases that do not involve AI.[1302] This factor may discourage them from seeking legal redress.

The main provisions of the AILD are, thus, designed to ensure that affected persons receive equivalent levels of protection in cases of harm caused by AI systems as they would under any other circumstances. Therefore, the AILD mandates that EU Member States revise their national civil liability frameworks in several respects.

1) Context: fault-based liability

The provisions of the AILD are intended to apply in the context of non-contractual fault-based liability claims. More specifically, the proposed Directive covers "claims for compensation of the damage caused by an output of an AI system or the failure of such a system to produce an output where such an output should have been produced."[1303] The claimants can be any legal or natural person who suffered damage from an AI system. The Directive adopts the definition of AI set out in the AI Act. The possible defendants are essentially providers or users (i.e., deployers) of AI systems.

Such wording is not easy to understand. It seems to follow that anything produced or generated by an AI system is covered. This includes, of course, anything produced by a generative AI model, if this output causes injury to natural or legal persons. The question arises as to whether this output must be contrary to the law. The answer to this question is not obvious. Content produced by generative AI can be harmful to people (e.g., incorrect information) without being illegal. In this context, it is conceivable that the provider will be liable if they have not complied with the "duty of care" standard referred to in the proposed Directive.

Whereas the notion and interpretation of fault are specific to each national law, the proposed AI Liability Directive specifies what constitutes fault, at least for the application of the presumption of causality it creates. Article 4(1) provides that the fault of the defendant consists in the noncompliance with "a duty of care laid down in Union or national law" and "directly intended to protect against the damage that occurred." The notion of "duty of care" is defined in Article 2(9) of the proposal as "the standard of conduct, set by national or Union law, in order to avoid damage to legal interests recognized at national or Union law level, including life, physical integrity, property and the protection of fundamental rights."

2) Creation of a rebuttable "presumption of causality"

The proposed directive offers a significant advantage to victims over traditional liability rules by creating a presumption that the defendant's negligence caused the harmful output of an AI system. Article 4(1)

[1301] "The 'black box' effect can make it difficult for the victim to prove fault and causality and there may be uncertainty as to how the courts will interpret and apply existing national liability rules in cases involving AI." *Id.* at 2.

[1302] *Id.* at 1

[1303] Artificial Intelligence Liability Directive (AILD), art. 2(5).

provides that national courts must presume the causal link between the fault of the defendant and the harmful output produced by the AI system (or the failure of the AI system to produce an output), when three conditions are met:

- The fault of the defendant (or someone for whom the defendant is responsible) has been established or presumed by the court (provided that fault consists of the noncompliance with a duty of care directly intended to protect against the damage that occurred).
- It is reasonably likely, based on the circumstances of the case, that the fault has influenced the output produced by the AI system or the failure of the AI system to produce an output.
- The claimant has demonstrated that the output produced by the AI system or the failure of the AI system to produce an output gave rise to the damage.

This general presumption is subject to adjustments and limitations depending on the context:

- If the claim is not related to a high-risk AI system, the presumption applies only where the court considers it excessively difficult for the claimant to prove the causal link.[1304]
- If the claim is related to a high-risk AI system, the presumption does not apply if the defendant demonstrates that sufficient evidence and expertise is reasonably accessible for the claimant to prove the causal link (in the case of a high-risk AI system).[1305]
- If the defendant used the AI system in the course of a personal, nonprofessional activity, the presumption applies only where the defendant materially interfered with the conditions of the operation of the AI system or if the defendant was required and able to determine the conditions of operation of the AI system and failed to do so.[1306]

Moreover, for claims related to high-risk AI systems, the proposed Directive precisely defines the conditions required in order to establish fault:

- If the defendant is a provider of a high-risk AI system, the claimant must demonstrate that the provider failed to comply with the obligations provided by the AI Act regarding high-risk AI systems.

- If the defendant is a user/deployer of a high-risk AI system, the claimant must prove that the user/deployer:
  - (a) did fail to use or monitor the AI system in accordance with the provided instructions of use or did not suspend or interrupt its use when necessary.
  - (b) exposed the AI system to irrelevant input data that they controlled and that was not fit for the system's intended purpose.

3) Disclosure of evidence related to High-Risk AI systems

[1304] *Id.* art. 4(5).
[1305] *Id.* art. 4(4).
[1306] *Id.* art. 4(6).

Article 3(1) of the proposed AI Liability Directive empowers national courts to mandate the release of relevant evidence regarding high-risk AI systems by the provider of the high-risk system (or someone who is subject to the same obligations under the AI Act) or a user (deployer). Such an injunction may be issued when a high-risk AI system is suspected to have caused damage, in the following circumstances:

- A potential claimant has previously asked the defendant to release the relevant evidence at its disposal but was met with a refusal, while being able to present facts and evidence sufficient to support the plausibility of a claim for damages.
- A claimant has brought a civil liability claim relating to this high-risk AI system.[1307]

It is worth noting that this injunctive power is potentially available for anticipated legal actions, as long as the claim is related to a high-risk AI system. However, the conditions are quite restrictive. The injunction may be granted exclusively if the claimant "has undertaken all proportionate attempts at gathering the relevant evidence from the defendant."[1308] Moreover, the courts' interpretation of "plausibility" remains to be seen, but there is a concern that the asymmetry of information between claimants and well-informed providers of high-risk AI systems—providers who are often international corporations hesitant to release liability-compromising information—could deter legal action. This is compounded by the technical complexity of the information, which is likely to be beyond the understanding of many lay people and perhaps even judges.

Courts must make sure that their orders for evidence disclosure are restrained to what is "necessary and proportionate to support a potential claim or a claim."[1309] They are empowered to order specific measures to preserve the evidence.[1310] They must consider the legitimate interests of all parties involved, such as the protection of trade secrets and intellectual property rights.[1311] In instances where revealing evidence would mean disclosing trade secrets or confidential information as defined by EU regulation, national courts will have the authority to implement measures to protect such confidentiality.

Lastly, if a defendant does not comply with the order to disclose evidence, the court shall presume "the defendant's non compliance with a relevant duty of care,"[1312] and, in particular, noncompliance with the requirements set out in the AI Act regarding high-risk AI systems. However, the presumption is rebuttable.

4) Conclusions on the proposed AI Liability Directive

The ambition of the proposed AI Liability Directive may seem relatively modest when juxtaposed with the extensive revisions being undertaken in the Product Liability Directive. Its scope is more specific, since the AI Liability Directive mainly targets high-risk AI systems. Moreover, the explanatory memorandum of the

[1307] *Id.* art. 2(6) AILD: "'claimant' means a person bringing a claim for damages that: (a) has been injured by an output of an AI system or by the failure of such a system to produce an output where such an output should have been produced; (b) has succeeded to or has been subrogated to the right of an injured person by virtue of law or contract; or (c) is acting on behalf of one or more injured persons, in accordance with Union or national law."
[1308] *Id.* art. 3(2).
[1309] *Id.* art. 3(4).
[1310] *Id.* art. 3(3).
[1311] *Id.* art. 3(4).
[1312] *Id.* art. 3(5).

proposal emphasizes that the harmonization intended by this new directive is "targeted" and does not aim to overhaul the general framework of civil liability established by the national laws of Member States.[1313] However, this limited harmonization could result in a significant divergence in legal decisions, considering that the concept of fault (negligence) is not uniformly defined and interpreted under Member States' national laws. In any case, individuals seeking compensation for harm caused by AI systems still have the option to use national laws if they offer more favorable provisions.[1314]

Furthermore, the proposal has been designed to tie in with the provisions of the AI Act, to which it systematically refers. While the progress of the AI Liability Directive is currently halted, it can be assumed that the future evolution of the text will take account of the significant changes made to the text of the AI Act during the negotiations. In particular, it is highly likely that the future drafts of the AI Liability Directive will include provisions on GPAI models and generative AI.

5.1.4. The Cyber Resilience Act

The European Union's Cyber Resilience Act (CRA)[1315] aims to enhance the security of digital products by establishing new safety standards within the EU. The European Parliament approved the CRA on March 12, 2024.[1316] Once the EU Council formally adopts it, the CRA will be published in the *Official Journal of the European Union* and, 20 days after publication, will enter into force. Most provisions of the CRA will become applicable 36 months after the Act's entry into force.

i) Scope of the Cyber Resilience Act

The Cyber Resilience Act regulates "products with digital elements" (PDEs), defined in Article 3(1) as a "software or hardware product and its remote data processing solutions." The concept of "remote data processing" solutions, as defined by Article 3(2) encompasses software that permits data processing from a remote location, is designed and developed by or on behalf of the PDE manufacturer, and is indispensable for the PDE to perform one of its functions, such as mobile apps for Internet of Things (IoT) products. For the CRA to apply to a specific PDE, it must be intended or reasonably foreseeable that the PDE will connect to a device or network for data transfer. This connection can be direct or indirect, virtual or physical (e.g., via software, wires, or radio signals). While the scope of the Act encompasses both software —including free and open-source software— and hardware products, it is further expanded by including "components" of PDEs, if they are placed on the EU market separately.

[1313] Proposal for an AI Liability Directive, *supra* note 1299, at recital 10. Recent EU laws have been adopted to impose cybersecurity standards. The Network Information Security Directive (NIS 2 Directive 2022/ 2555) targets critical infrastructure, while the Digital Operational Resilience Act (Regulation 2022/2554) focuses on the financial services sector. The Cyber Resilience Act (CRA) takes a more horizontal approach.

[1314] *Id.* at Recital 14.

[1315] European Parliament legislative resolution P9_TA(2024)0130 of 12 March 2024 on the proposal for a regulation on horizontal cybersecurity requirements for products with digital elements (**Cyber Resilience Act**), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0130_EN.pdf

[1316] There are two exceptions: manufacturers' reporting obligations will apply 21 months after the CRA's entry into force; provisions concerning the notification of conformity assessment bodies will apply 18 months after the CRA's entry into force.

Therefore, the CRA applies to nearly all digital products, including end devices, such as laptops, smartphones, and desktops; networking appliances; IoT devices; and software. It also covers CPUs, video cards, and both open and closed-source software libraries. Given this broad scope, most AI systems fall under the scope of the Cyber Resilience Act. The Act does not apply to products governed by sector-specific legislation, such as medical devices, provided that the sector-specific regulations adequately address cybersecurity risks and requirements.[1317] Additionally, the Act excludes cloud solutions that do not qualify as remote data processing solutions, including Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).[1318]

The Cyber Resilience Act does not grant a general exemption for open-source software, defined in Article 3(48) as "software the source code of which is openly shared and and which is made available under a free and open-source licence which provides for all rights to make it freely accessible, usable, modifiable and redistributable". However, open-source softwares developed outside of commercial activity are not covered by the Regulation, which only covers free and open-source software "made available on the market and therefore supplied for distribution or use in the course of a commercial activity".[1319] The original draft of the Act raised concerns that the term "commercial" would be interpreted broadly, potentially subjecting all acts of publishing and sharing software and source code to the risk of fines for developers. However, the final version clarifies that commercial activity requires that softwares are monetized by their manufacturers.

The Act also introduces the category of "open-source stewards." Article 3(14) of the Act provides that this category includes individuals or organizations that provide support for the development of specific products with digital elements that qualify as free and open-source software intended for commercial activities and ensure the viability of those products. These stewards must comply with obligations such as establishing cybersecurity policies, encouraging the responsible disclosure of vulnerabilities, and collaborating with authorities to address security risks.

ii) Obligations provided by the Cyber Resilience Act

The Cyber Resilience Act mandates that regulated entities minimize and address cybersecurity risks throughout the entire PDE lifecycle, imposing the strictest obligations on manufacturers. This category broadly encompasses anyone who develops or manufactures PDEs, or has such products designed, developed, or manufactured, and markets them under their name or trademark. This includes operators who

[1317] These exclusions include medical devices, motor vehicles, military hardware, certified aviation products, marine equipment, spare parts that replace identical components, and digital elements developed or modified exclusively for national security or defense purposes.

[1318] Which are regulated by the NIS 2 Directive, *see* Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union (**Network and Information Security 2 - NIS 2 Directive**), http://data.europa.eu/eli/dir/2022/2555/2022-12-27.

[1319] Cyber Resilience Act, Recital 18; *See also* Paul Sawers, *Open source foundations unite on common standards for EU's Cyber Resilience Act,* TechCrunch (April 2, 2024), https://techcrunch.com/2024/04/02/open-source-foundations-unite-on-common-standards-for-eus-cybersecurity-resilience-act/; Ashwin Ramaswami & Mirko Boehm, *Understanding the Cyber Resilience Act: What Everyone involved in Open Source Development Should Know*, Linux Foundation, (September 8, 2023) https://www.linuxfoundation.org/blog/understanding-the-cyber-resilience-act

substantially modify the product. The CRA also extends its reach to the entire supply chain, encompassing importers and distributors. The regulation applies regardless of where these entities are located, as long as their products are made available on the EU market.

The Cyber Resilience Act (CRA) outlines the cybersecurity and other requirements applicable to all Products with Digital Elements (PDEs) in Articles 13, 14, and Annex I. These obligations include a combination of product requirements, information obligations, and the adoption of internal processes. Within this framework, the CRA establishes the following:
1. Rules for placing products with digital elements on the market to ensure their cybersecurity.
2. Essential requirements for the design, development, and production of PDEs.
3.Essential requirements for manufacturers' vulnerability handling processes to ensure cybersecurity throughout the product lifecycle.
4. Rules on market surveillance and enforcement.

In particular, the CRA imposes four primary categories of obligations on manufacturers: conducting risk assessments, maintaining documentation, performing conformity assessments, and reporting vulnerabilities. Manufacturers must perform conformity assessments before commercialization to ensure that the digital products they market comply with "essential cybersecurity requirements" ("ECRs"). These essential requirements include:

- ensuring products are free of known vulnerabilities,
- implementing secure settings and access controls,
- protecting data confidentiality, integrity, and availability,
- limiting data processing and attack surfaces,
- mitigating exploitation risks, and
- providing security logs.[1320]

While the requirements outlined in the CRA apply to all PDEs, the process for certifying compliance with these measures varies according to the product classification established by the Act. Most manufacturers can self-assess, but manufacturers of "Important" and "Critical" PDEs must undergo third-party conformity assessment. Under Article 7, Important PDEs are those that either primarily perform functions critical to the cybersecurity of other products, networks, or services or that carry a significant risk of causing substantial adverse effects, such as disruption, control, or damage to a large number of other products or to the health, security, or safety of their users. This category includes items such as firewalls, password managers, antivirus software, or identity management systems, as listed in Annex III. According to Recital 46, the "Critical PDEs" covered by Article 8 "have a cybersecurity-related functionality and perform a function which carries a significant risk of adverse effects in terms of its intensity and ability to disrupt, control or damage a large number of other products with digital elements through direct manipulation." Examples of Critical PDEs include hardware devices with security boxes, smart meter gateways within smart metering systems, and other advanced security devices, such as secure crypto processing units and smartcards, as outlined in Annex IV.

[1320] Cyber Resilience Act, Annex I.

Manufacturers must ensure their PDEs comply with the CRA before they can put them on the market in the EU. They must draw up a "declaration of conformity" to provide information regarding the compliance of their products to the essential requirements specified in the Act. Distributors and importers are required to verify that the manufacturer has completed the necessary assessments before placing the product on the EU market. The CRA provides a presumption of conformity with essential requirements if the PDE complies with harmonized technical standards at the EU level.

Finally, manufacturers are obligated to maintain the security of their products throughout their lifecycle. They are required to manage product vulnerabilities effectively through regular testing, patch management, responsible disclosure programs, and clear documentation. They must report any actively exploited vulnerabilities and severe security incidents affecting their products to the competent authorities.[1321]

iii) Enforcement of the Cyber Resilience Act

Each Member State must designate one or more market surveillance authorities to enforce the Cyber Resilience Act at the national level. For PDEs classified as "high-risk AI systems" under the AI Act, the same national market surveillance authority will be responsible under both the CRA and the AI Act. The CRA establishes an Administrative Cooperation (AdCo) group at the EU level, comprising all national market surveillance authorities and representatives from the EU Commission.

EU Member States are competent to lay down the level of penalties for non-compliance with the CRA –but in relation to infringements against the "essential cybersecurity requirements," the CRA introduces a sanction regime for noncompliance. Potential maximum fines range from €5 million to €15 million or 1% to 2.5% of global annual turnover, whichever is greater.

Overall, the cybersecurity requirements in the Cyber Resilience Act complement and intensify the obligations imposed on AI system providers by the AI Act. For instance, high-risk AI systems that meet the essential requirements of the CRA should also be considered compliant with the cybersecurity requirements set forth in the AI Act, as provided by Article 12 of the Cyber Resilience Act.

KEY TAKEAWAYS

- The emergence of generative AI in Europe has taken place within an established legal framework that heavily regulates technology companies. This framework encompasses various EU statutes, including the General Data Protection Regulation (GDPR), the New Copyright Directive, and the Digital Services Act (DSA). Nevertheless, these existing laws have been deemed inadequate for regulating the deployment of artificial intelligence in the EU. As a result, the EU has recently adopted additional legislative measures, such as the AI Act, the Product Liability Directive, and the Cyber Resilience Act. The prospective adoption of the AI Liability Directive could further harmonize the liability framework within the EU.

[1321] Which are the Computer Security Incident Response Team (CSIRT) and the EU Agency for Cybersecurity (ENISA).

- The recently adopted AI Act represents the most comprehensive framework for regulating artificial intelligence to date. This Act imposes varying obligations on AI systems based on their potential risk to health, safety, and fundamental rights. Initially, the AI Act aimed to address risks associated with specific use cases, particularly in sensitive sectors. The Act categorizes risks based on the "intended" use of AI systems and classifies them into four risk categories: unacceptable risk, which warrants prohibition; high risk, which necessitates stringent obligations; limited or "transparency" risk, which requires disclosure of information to users; and minimal or no risk. During the negotiation process, provisions were added to regulate general-purpose (i.e., foundation) models, shifting the focus from specific use cases to the technology itself. Consequently, the AI Act now includes provisions governing general-purpose AI models, with more rigorous obligations imposed on those models identified as having "systemic risk," essentially the most capable models.

- Within this framework, providers of generative AI models and systems will be subject to two distinct categories of regulations. First, they must supply information and documentation to stakeholders, adopt a copyright policy, and maintain transparency about their training data sets. If their model poses a systemic risk, these obligations become even more stringent and include providing additional information and complying with requirements related to cybersecurity, red teaming, risk mitigation, incident reporting, and model evaluation. Second, providers who supply AI systems built on general-purpose AI models, such as a generative chatbot running on a foundation model, face an additional set of rules. These include transparency obligations specific to chatbots and generative AI systems, such as informing users they are interacting with an AI system and implementing watermarking. Finally, if the AI system is used in a sensitive sector, such as healthcare or the judicial system, it may be categorized as high risk, thereby subjecting providers to the stringent obligations applicable to high-risk AI systems.

- While existing EU frameworks often reflect traditional regulatory methods, the AI Act incorporates elements of co-regulation. For instance, its implementation involves various stakeholders, such as the European Artificial Intelligence Board, made up of one representative from each Member State; the advisory forum, which represents a diverse group of stakeholders, including industry representatives, startups, small- and medium-sized enterprises (SMEs), civil society, and academia; and the scientific panel of independent experts. The development of codes of practice, for example, will be a collaborative effort involving industry representatives, the scientific community, civil society, and the AI Office. Adherence to these codes will be essential for providers of general-purpose AI models to establish compliance. However, despite the AI Act's emphasis on dialogue with industry and stakeholders, it does not specifically assign a role to researchers and academic institutions, unlike the provisions in the DSA.

- Implementing the recently adopted EU laws governing AI will pose significant challenges. Although the AI Act, as a Regulation, is directly applicable in EU Member States, effective enforcement will necessitate efficient collaboration between the EU and Member State authorities. The involvement of numerous bodies and stakeholders in the implementation process could also substantially slow its application. Furthermore, the EU Commission will need to adopt numerous delegated acts, implementing acts, and application measures to ensure the effective implementation

of the Act. The risk of uneven enforcement is even more pronounced with the Product Liability Directive, which requires precise transposition into national laws. As a result, disparities among Member States may persist, particularly in areas such as civil liability.

- Like the GDPR, the AI Act is likely to have a far-reaching impact beyond the European Union. Specifically, the AI Act will have significant extraterritorial reach, applying to all AI services offered and used within the EU. Consequently, the AI Act is likely to set global standards, as it would be impractical for companies to create separate AI models and systems for different markets.

- The AI Act is complemented by the simultaneously adopted Cyber Resilience Act (CRA). The CRA imposes several cybersecurity rules on internet-connected "products with digital elements," encompassing both hardware and software products, including AI systems. The CRA aims to enhance the security of digital products and services throughout their lifecycle. It introduces obligations for manufacturers, importers, and distributors to ensure that products with digital elements meet specified cybersecurity standards. Key provisions include requirements for manufacturers to effectively manage product vulnerabilities through regular testing, patch management, responsible disclosure programs, and clear documentation.

- While comprehensive, the current EU legislation applicable to AI does not guarantee absolute legal certainty. The GDPR exemplifies this issue. Drafted before the emergence of generative AI systems, the GDPR was not designed to accommodate the unique characteristics of these technologies. Ensuring that generative AI companies effectively comply with the GDPR is currently very challenging: Establishing the legal basis for processing training data, ensuring the accuracy of AI-generated output, and guaranteeing the right to rectification for data subjects are all complex tasks. These difficulties have subjected AI companies to investigations and legal claims regarding the compliance of their generative AI tools with the GDPR.

- From a copyright law perspective, EU Member States have already implemented EU Directives applicable to developers of generative AI, notably the New Copyright Directive 2019/790. This directive includes the so-called “Text and Data Mining exception,” which allows the collection and use of web-scraped data unless explicitly reserved by copyright holders. However, several copyright issues remain unresolved. Determining whether merely composing a prompt entitles the user of an AI model to claim copyright over the content generated in response to that prompt presents a significant challenge. Additionally, there is uncertainty over whether AI-generated content can constitute an infringement of existing copyrights.

- The issue of liability for AI systems, particularly concerning AI-generated outputs, remains uncertain. While the Digital Services Act (DSA) exempts hosting providers from liability, this exemption does not directly apply to providers of generative AI tools. Currently, liability questions concerning developers, providers, or users of generative AI tools are governed by the national laws of EU Member States, which typically apply a traditional fault-based liability principle. These laws are expected to evolve with the recent adoption of the new Product Liability Directive, which classifies software, including AI systems, as products that can be defective, thereby holding their producers liable. However, the implementation of the Product Liability Directive necessitates

precise transposition into national laws, creating a risk of uneven application of these provisions across the EU. Additionally, the AI Liability Directive proposed by the EU Commission, though currently under negotiation, may eventually be adopted and provide presumptions of fault and causality to facilitate liability claims against AI system providers and operators.

5.2. China

While some leadings models, including OpenAI's ChatGPT and Google's Gemini, are not available in China,[1322] open models, such as Meta's Llama 2 and Llama 3, are embraced by Chinese tech firms. Alibaba Group Holding and Baidu have added support for Meta's large language model to their cloud computing platforms.[1323] Additionally, a large number of Chinese companies and research institutes have built their own generative AI models. Among them are Alibaba's Tongyi Qianwen, Tencent's Hunyuan, Huawei's Pangu, and Beijing Academy of AI's Aquila.[1324] In March 2023, Baidu launched Ernie Bot, the well-known technology giant's own ChatGPT-equivalent chatbot.[1325]

Regulation and innovation are frequently perceived as opposing forces in the discourse on regulatory frameworks for emerging technologies. However, in China, these forces appear to converge through comprehensive legislative measures aimed at governing AI technologies. China is notably one of the leading nations with the most extensive AI legislation, particularly concerning generative AI.

This section begins with an overview of China's general AI strategy. Next, it examines the legal frameworks for data and copyright protection, followed by a review of the successive legislations adopted by the Chinese government to regulate AI. Finally, the section concludes with a discussion of the ethical frameworks developed in China concerning AI.

5.2.1. General overview of China's AI strategy

[1322] Major AI models, such as ChatGPT and Gemini, have excluded Mainland China and Hong Kong in their official lists of supported countries. ChatGPT was already blocked in China by the Chinese firewall, but developers were previously able to use virtual private networks to access OpenAI's tools for fine-tuning their generative AI applications and benchmarking their research. Recently, however, OpenAI informed its users in China that they would be blocked from using its tools and services starting from July 9, 2024. *ChatGPT Supported Countries*, OPENAI, https://help.openai.com/en/articles/7947663-chatgpt-supported-countries (last visited May 4, 2024); *Where you can use the Gemini web app with Google's 1.0 Pro model*, GOOGLE, https://support.google.com/gemini/answer/13575153?hl=en (last visited May 4, 2024); *OpenAI Taking Steps to Block China's Access to Its AI Tools*, BLOOMBERG (June 25, 2024), https://www.bloomberg.com/news/articles/2024-06-25/openai-warns-it-will-block-access-to-ai-tools-from-china?leadSource=uverify%20wall.

[1323] *Alibaba, Baidu rush to add support for Meta's Llama 3 on their cloud computing platforms*, SOUTH CHINA MORNING POST (Apr. 23, 2024), https://www.scmp.com/tech/tech-trends/article/3259945/alibaba-baidu-rush-add-support-metas-llama-3-their-cloud-computing-platforms; Amy Hawkins, *Chinese developers scramble as OpenAI blocks access in China*, THE GUARDIAN, (July 9, 2024), https://www.theguardian.com/world/article/2024/jul/09/chinese-developers-openai-blocks-access-in-china-artificial-intelligence

[1324] Paul Triolo & Anarkalee Perera, *This is the state of generative AI in China*, THE CHINA PROJECT (Sept. 20, 2023), https://thechinaproject.com/2023/09/20/this-is-the-state-of-generative-ai-in-china/.

[1325] Zayi Yang, *Chinese tech giant Baidu just released its answer to ChatGPT*, MIT TECH. REV. (Mar. 16, 2023), https://www.technologyreview.com/2023/03/16/1069919/baidu-ernie-bot-chatgpt-launch/.

China is unique in both its conservative domestic goals prioritizing political stability and its ambitions to assume global leadership in technology development and governance.[1326] All this plays out in front of its strategic competition with the US in the international arena.[1327] Collectively, these characteristics shape China's legislative approach that stands apart from both the EU and the US. The Chinese government strives to exercise control over both the development of AI technologies and their use. And, it attempts to continuously empower its technological capabilities while reducing reliance on foreign supply chains and key technologies.[1328]

China's overarching AI strategy was articulated in a plan formulated in 2017. More recently, the country has sought to assert its influence on the global stage by adopting a clear stance on AI governance, with the Global AI Governance Initiative of China.

5.2.1.A. The New Generation AI Development Plan (2017)

In July 2017, China's State Council issued a notice on the New Generation AI Development Plan (Development Plan), marking China's first systematic policy document in artificial intelligence.[1329] The Development Plan aims to establish China as a world leader in AI by 2030, with milestones set for 2020 and 2025.[1330] The strategy includes significant investment in AI research and development, promoting the integration of AI across various sectors, and developing comprehensive regulations and ethical guidelines for AI technologies.

Unfolding in three stages, the Development Plan outlines strategic objectives, critical tasks, resource allocation, and measures for AI development:

[1326] Dominic Paulger, *Navigating Governance Frameworks for Generative AI Systems in the Asia-Pacific*, FUTURE OF PRIVACY FORUM, (May 2024), https://fpf.org/wp-content/uploads/2024/05/Navigating-Governance-Frameworks-for-Gen-AI-Systems-in-the-Asia-Pacific.pdf.

[1327] Ashyana-Jasmine Kachra, *Making Sense of China's AI Regulations*, HOLISTIC AI (Feb. 12, 2024), https://www.holisticai.com/blog/china-ai-regulation#:~:text=The%20rules%20seek%20to%20balanceand%20other%20violence%20or%20misinformation; Matt Sheehan, *China's AI Regulations and How They Get Made*, CARNEGIE ENDOWMENT FOR INT'L PEACE (July 10, 2023), https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

[1328] Kachra, *Making Sense of China's AI Regulations*, *Id.*

[1329] Guanyu Yinfa "Xinyidai Rengong Zhineng Fazhan Guihua" de Tongzhi (关于印发《新一代人工智能发展规划》的通知) [Notice of Issuing the New Generation Artificial Intelligence Development Plan] (promulgated by the State Council, effective July 8, 2017), https://www.gov.cn/zhengce/zhengceku/2017-07/20/content_5211996.htm; *translated in* Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017), DIGICHINA (Aug. 1, 2017), https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/; Guihua dui Rengong Zhineng Weilai Fazhan Jinxing le Xitong Bushu (《规划》对人工智能未来发展进行了系统部署) [The Plan provides a systematic deployment of the future development of AI] (Sept. 24, 2017), https://www.most.gov.cn/xwzx/twzb/fbh17072101/twzbzbzy/201707/t20170724_134186.html.

[1330] Huw Roberts et al., *The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation*, 36 AI & SOC. 59 (2021), https://link.springer.com/article/10.1007/s00146-020-00992-2.

- By 2020, China aimed for its AI technologies and applications to reach global advanced standards, transforming the AI sector into a key driver of economic growth. Furthermore, AI applications were expected to provide new opportunities for enhancing public welfare.
- By 2025, China's goal is to achieve substantial breakthroughs in AI foundational theories, positioning certain technologies and applications at the forefront of global innovation. AI is also expected to drive a transformative change in China's industrialization and overall economy, advancing the formation of an intelligent society.
- By 2030, China's vision is for its AI theories, technologies, and applications to collectively achieve a leading position on the global stage, making China a major hub for AI innovation. This achievement would mark a significant milestone in establishing an intelligent economy and society.

The New Generation AI Development Plan remains a cornerstone of the country's AI strategy. The introduction of new regulations and initiatives in 2023 and 2024 align with the strategic goals outlined in the 2017 plan. These measures underscore China's commitment to addressing AI safety, security, and ethical considerations while fostering innovation and technological advancement.

5.2.1.B. The *Global AI Governance Initiative of China* (2023)

Internationally, the *Global AI Governance Initiative of China* (the "*Initiative*"), introduced by President Xi Jinping in October 2023 during the third Belt and Road Forum for International Cooperation, aims to shape the development and governance of artificial intelligence on a global scale.[1331] This initiative represents China's strategic move to assert its influence over the global governance of artificial intelligence. The *Initiative* is notable for its emphasis on international collaboration and equitable governance of AI technologies, advocating for a balanced approach that considers both opportunities and risks. The *Initiative* opposes technological monopolies and promotes global cooperation to prevent the misuse of AI technologies. It also highlights the need for developing countries to have a significant voice in global AI governance.

Key features of the *Initiative* include a focus on inclusivity and equity, ensuring that all countries, regardless of economic status or political systems, can participate in AI development and governance. Another key feature makes a commitment to developing ethical norms, privacy protections, and legal structures, aligning with global calls for robust regulations. And it expresses an emphasis on sustainable development and human-centric principles, which prioritize human welfare and security.

5.2.2. Data and copyright protection

[1331] Wang Cong & Yin Yeping, *China launches Global AI Governance Initiative, offering an open approach in contrast to US blockade*, GLOB. TIMES (Oct. 18, 2023), https://www.globaltimes.cn/page/202310/1300092.shtml; Quanqiu Rengong Zhineng Zhili Changyi (全球人工智能治理倡议) [Global Artificial Intelligence Governance Initiative], CYBERSPACE ADMIN. OF CHINA (Oct. 18, 2023), https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm; for a translated version, see *Global Artificial Intelligence Governance Initiative*, EMBASSY OF CHINA IN THE KINGDOM OF SAUDI ARABIA (Oct. 24, 2023), http://gd.china-embassy.gov.cn/eng/zxhd_1/202310/t20231024_11167412.htm.

Developers of generative AI models in China, like their counterparts elsewhere, must comply with legislation designed to protect personal data and intellectual property rights.

5.2.2.A. Data protection

The Chinese government enacted the Personal Information Protection Law (PIPL) in 2021.[1332] It is a comprehensive data privacy legislation that closely resembles the EU's GDPR and aims to protect Chinese citizens' personal and sensitive data by regulating access, processing, and sharing of such information. The PIPL also imposes restrictions on data-hungry AI companies, requiring informed consent for various data-processing activities. It grants individuals significant data rights, such as the rights to amend, delete, and request copies of the information collected about them.[1333]

Article 13(6) of the PIPL allows data controllers to process publicly available personal data to a reasonable extent without the data subject's consent if the data subject personally disclosed the data; or the data were otherwise legally disclosed. However, consent is required if the data subject explicitly objects to the processing or if the processing could significantly impact an individual's rights and interests.[1334] In such cases, the data controller must obtain the data subject's consent, which is valid only if it is voluntarily given, explicit, and fully informed.[1335] Additionally, consent is necessary for the processing of sensitive personal data.[1336]

The data controller must also inform data subjects about the necessity of processing their data and the potential impact on their rights and interests.[1337] Article 8 of the PIPL mandates that data controllers ensure the accuracy and completeness of personal data to prevent any adverse effects on individuals' rights and interests resulting from inaccurate or incomplete data. Finally, the PIPL, similar to the GDPR, explicitly prohibits the use of automated decision-making systems that result in discriminatory treatment of consumers, such as differential pricing based on personal data.

5.2.2.B. Copyright protection

In China, numerous legal claims have been brought before the courts over the past several years concerning the application of copyright law to developers of generative AI. The issues raised in the courts concern both the copyrightability of AI-generated works and the potential for these works to infringe upon existing copyrighted material.

1) Copyrightability of AI-generated work

[1332] Personal Information Protection Law (promulgated by the Standing Comm. Nat'l People's Cong., Aug. 20, 2021, rev'd Sept. 7, 2021, effective Nov. 1, 2021), 2021 P.R.C. LAWS, https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/.

[1333] *Id.*

[1334] *Id.* art. 27, Personal Information Protection Law (PIPL).

[1335] *Id.* art. 14.

[1336] *Id.* art. 29.

[1337] *Id.* art. 30.

As early as 2018, the question of whether AI-generated works qualify for copyright protection has been examined by numerous Chinese courts.

a) *Beijing Film Law Firm v. Beijing Baidu Netcom Science & Technology Co. Ltd*. (2018)

Ruling in China's first case involving copyright protection, the Beijing Internet Court found that AI-generated graphs lacked "originality" and did not qualify for copyright protection. However, the court recognized the software user's rights and interests in the generated report without attributing authorship. In this case, the copyright infringement lawsuit concerned graphs produced by a legal statistical data analysis software and were subsequently published in an article by the plaintiff. The court ruled that these graphs, despite being derived from collected data and processed through software, lacked *originality* because they simply reflected data changes and would produce identical outputs under similar conditions. Consequently, these graphs did not meet the criteria for protection as graphic works under copyright law, leading to the dismissal of the plaintiff's copyright claim.

In the same ruling, the court also examined the nature and ownership rights of an analysis report automatically generated by the legal statistical data analysis software (i.e., Wolters Kluwer Database). It found that, while the report exhibited originality and adhered to the formal aspects of written works, it could not be classified as a "work" under copyright law because it was not created by a *natural person*.[1338] The contributions from software developers and users did not fulfill the necessary threshold for original expression to qualify for authorship. However, the court noted that, although the report could not be copyrighted, the software user, through their specific inputs and use of the software, should possess rights and interests in the report to facilitate its utilization and dissemination, thereby recognizing their involvement without attributing authorship.

b) *Shenzhen Tencent v. Shanghai Yingxu* (2019)

In this case, the Shenzhen Nanshan District People's Court granted copyright protection to an AI-assisted article, acknowledging the human intellectual activities involved in its creation.[1339] A Chinese company published an article in 2019 on its website that had been generated using a set of data and algorithm-based intelligent writing assistance software. When another Chinese company reprinted the article on its website without the permission of the author company, the author company sued.

In this case, Nanshan District People's Court in Shenzhen held that the author company could claim copyright as the article qualified as a "written work" under Chinese copyright law despite being generated by software. The court held that the efforts of the company's creative team in terms of data input, template,

---

[1338] Beijing Film Law Firm v. Beijing Baidu Netcom Science & Technology Co Ltd., (2018) Jing 0491 Min Chu No. 239, BEIJING INTERNET CT. (Apr. 25, 2019), http://www.chinadaily.com.cn/specials/BeijingInternetCourtCivilJudgment(2018)Jing0491MinChuNo.239.pdf; Beijing Internet Court Civil Judgment (2018) Jing 0491 Min Chu No 239, BEIJING INTERNET CT. (May 28, 2019), https://english.bjinternetcourt.gov.cn/2019-05/28/c_168.htm.

[1339] Shenzhen Tencent v. Shanghai Yingxun, (2019) Yue 0305 Min Chu No. 14010 Civil Judgment, NANSHAN DISTRICT PEOPLE'S CT. SHENZHEN, GUANGDONG PROVINCE (Dec. 24, 2019); Zhou Bo, *Artificial Intelligence and Copyright Protection- Judicial Practice in Chinese Courts*, WIPO (2020), https://www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/conversation_ip_ai/pdf/ms_china_1_en.pdf.

and corpus style choices qualified as intellectual activities. The article was the result of human intellectual activity assisted by AI and could not have been said to be autonomously generated by AI. Thus, by granting protection to the author company, the court did not deviate from the general rule that the work must result from the author's intellectual creation.[1340]

c) *Li Yunkai v. Liu Yuanchun* (2023)

The Beijing Internet Court (BIC) affirmed the copyrightability of an *AI-generated image*, based on the plaintiff's intellectual contributions in prompt selection and parameter adjustment. In November 2023, the BIC upheld a person's claim to copyright ownership in an AI-generated image.[1341] The case involved a plaintiff using Stable Diffusion, a popular open-source text-to-image generative AI model, to generate an image of a woman. The defendant used the same image on her blog without the plaintiff's permission, and the plaintiff sued for copyright infringement. The court ruled in favor of the plaintiff, directing the defendant to publicly apologize and pay the plaintiff 500 Yuan (USD $70) in damages and 50 Yuan (USD $7) in court fees.[1342]

Among other grounds, the BIC held that the Chinese law requirement that the image be a result of the plaintiff's "intellectual achievement" had been met in this case. The BIC noted that the plaintiff chose more than 150 prompts, including negative prompts,[1343] organized in a specific order, and established certain parameters. He persistently tweaked and altered these prompts and parameters until the final image matched his vision. These actions, said the court, were sufficient to demonstrate that the contested image was produced due to the plaintiff's intellectual contributions.

Though the 2023 BIC judgment is not precedential under Chinese judicial practice,[1344] the decision was likely influenced by the provisions of the *Interim Administrative Measures for Generative Artificial Intelligence Services (see below section 5.2.3.C)*. They require government authorities to balance the

[1340] Shenzhen Tencent v. Shanghai Yingxun, (2019) Yue 0305 Min Chu No. 14010 Civil Judgment, NANSHAN DISTRICT PEOPLE'S CT. SHENZHEN, GUANGDONG PROVINCE (Dec. 24, 2019); Zhou Bo, *Artificial Intelligence and Copyright Protection- Judicial Practice in Chinese Courts*, WIPO (2020), https://www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/conversation_ip_ai/pdf/ms_china_1_en.pdf.

[1341] Yuqian Wang & Jessie Zhang, *Beijing Internet Court Grants Copyright to AI-Generated Image for the First Time***,** KLUWER COPYRIGHT BLOG (Feb. 2, 2024), https://copyrightblog.kluweriplaw.com/2024/02/02/beijing-internet-court-grants-copyright-to-ai-generated-image-for-the-first-time/.

[1342] Keith Kelly, *Computer Love: Beijing Court Finds AI-Generated Image is Copyrightable in Split with United States*, NAT'L LAW REV. (Dec. 4, 2023), https://www.natlawreview.com/article/computer-love-beijing-court- finds-ai-generated-image-copyrightable-split-united.

[1343] Yuqian Wang & Jessie Zhang, *see supra* note 1341.

[1344] Bruce Wang et al., *Beijing Internet Court grants copyright protection for AI artworks, but copyrightability debate of AI-generated output continues*, HOGAN LOVELLS (Dec. 6, 2023), https://www.engage.hoganlovells.com/knowledgeservices/news/beijing-internet-court-grants-copyright-protection-for-ai-artworks-but-copyrightability-debate-of-ai-generated-output-continues.

promotion of innovation with governance and encourage innovative applications of generative AI in all industries.[1345]

d) Copyright protection for the benefit of the user of generative AI

From the above cases, it would seem that Chinese courts have taken a different approach to that taken in the US, where the U.S. Copyright Office considers that prompting alone generally is *not* a sufficiently human contribution to be regarded as establishing authorship of the resulting output (*see below section 5.3.1.B.*).[1346] While Chinese courts do not grant intellectual property rights for purely AI-generated works that lack originality, they now recognize copyright for users who create original works with the assistance of generative AI. The courts justify this protection by noting that users contribute to the creation process through their "prompts."

2) Can AI-generated content infringe on intellectual property rights?

Chinese courts readily punish generative AI developers when their tools produce outputs that closely resemble or duplicate copyrighted content. Notably, a recent decision has recognized the protection of an individual's voice under personality rights.

a) The *Ultraman* case

In February 2024, the Guangzhou Internet Court declared a generative AI service provider liable for enabling the generation of copyright-infringing content through its website.[1347] The lawsuit began after a Japanese production company that owns the copyright to the famous animated cartoon series *Ultraman* gave a Chinese company an exclusive license for that character (including the right of reproduction, the right to prepare derivative works, and the right to enforce). The Chinese company, upon inputting some prompts into a generative AI website with a text-to-image generation function, discovered that the website reproduced exact images of the *Ultraman* character. Thus, the Chinese company sued the generative AI company for copyright infringement.

The Guangzhou Internet Court found the generative AI company liable for copyright infringement. In addition to holding that the defendant's AI website had violated the plaintiff's exclusive right to reproduce the Ultraman character, the Court also held that the defendant violated the plaintiff's exclusive right to

---

[1345] art. 5, *Interim Administrative Measures for Generative Artificial Intelligence Services*, CHINA L. TRANSLATE (July 13, 2023), HTTPS://WWW.CHINALAWTRANSLATE.COM/EN/GENERATIVE-AI-INTERIM/; Angela Huyue Zhang, *China's Short-Sighted AI Regulation*, PROJECT SYNDICATE (Dec. 8, 2023), https://www.project-syndicate.org/commentary/risks-of-beijing-internet-court-ruling-allowing-copyright-of-ai-generated-content-by-angela-huyue-zhang-2023-12.

[1346] Indeed, if generative AI shifts the locus of creativity toward devising prompts and away from crafting the expressive work itself, this could put significant strain on long-standing legal doctrines and incentive structures that underpin US copyright law. Mark A. Lemley, *How Generative AI Will Turn Copyright on its Head*, COLUM. SCI. & TECH. L. REV. (forthcoming 2024), https://ssrn.com/abstract=4517702 or http://dx.doi.org/10.2139/ssrn.4517702.

[1347] Christine Yiu & Toby Bond, *Liability of AI Service Providers for Copyright Infringement: Guangzhou Internet Court reaches world's first decision*, BIRD & BIRD (Apr. 10, 2024), https://www.twobirds.com/en/insights/2024/china/liability-of-ai-service-providers-for-copyright-infringement.

prepare derivative works of the licensed work. This was because the plaintiff demonstrated that the images generated from prompts asking for modified versions of the protected character (such as "Ultraman with long hair" or "Ultraman in cartoon style") bore a substantial resemblance to the original licensed work.[1348] The court directed the defendant to pay a compensation of 10,000 Yuan (~USD 1,400), though the plaintiff had sought compensation of 300,000 Yuan (~USD 41,500).

Another important part of this judgment is the discussion of the defendant company's duty of care. This part is discussed in more detail in section 5.2.3.C(3) below.

b) The AI-generated voice application case

The Beijing Internet Court also dealt with a case, in April 2024, involving the use of a person's voice in training an AI text-to-speech model to generate similar sounding voice outputs.[1349] The case was filed by a voice actor who discovered some voice dubbing works being circulated on popular applications used her voice. The voice actor had a voice recording contract with a media company that owned the copyright in the sound recordings made from her voice. Without the voice actor's permission, the media company licensed the sound recordings to an AI software development company to use for commercial and noncommercial purposes. This AI software company then used the voice recordings to train an AI text-to-speech application which was later used by another company that operated a voice dubbing app.[1350]

The BIC held that, under the Chinese Civil Code, a person's voice rights are protected as a special personal right. The court ruled that the voice actor's rights in this situation would apply to the AI-generated voice outputs, as the public could identify the voice actor's speech in the AI-generated outputs. An important takeaway from this case is the court's ruling on assignment of liability. The court held that both the media company and the AI software development company were liable for using the sound recording without the voice actor's consents. The fact that the media company owned the copyright to the sound recording did not mean that it could authorize third parties to use the plaintiff's voice in the sound recording to train AI models to generate similar sounding voice outputs. The court ordered both companies to pay the voice actor 250,000 Yuan (~USD 34,500) as compensation for her loss. The court did not attribute any liability to the company operating the voice dubbing app.[1351]

5.2.3 China's regulatory initiatives on AI technologies

[1348] Seagull Song & Wang Mo, *China's First Case on AIGC Output Infringement- Ultraman*, KING & WOOD MALLESONS (Feb. 28, 2024), https://www.kwm.com/global/en/insights/latest-thinking/china-s-first-case-on-aigc-output-infringement-ultraman.html.

[1349] Seagull Song & Wang Mo, *China's First Case regarding AI-generated Voice Infringement*, KING & WOOD MALLESONS (Apr. 26, 2024), https://www.kwm.com/cn/en/insights/latest-thinking/china-s-first-case-regarding-ai-generated-voice-infringement.html.

[1350] *Id.*

[1351] *Id.*

China's pursuit of technological self-reliance aligns with its legislative shift from an application-specific, adaptive approach to a more comprehensive legal framework for AI technologies.[1352] China's centralized approach to regulating AI currently involves two key entities: the Cyberspace Administration of China (CAC), which issues regulations, and the National Information Security Standardization Technical Committee (TC260), which develops technical standards.

China's governance and regulatory initiatives regarding AI started in 2019. Initially, the government issued non-binding guidance. In June 2019, the National New Generation Artificial Intelligence Governance Expert Committee, established by the Ministry of Science and Technology, issued eight principles under the *Governance Principles for New Generation AI: Develop Responsible Artificial Intelligence* ("*Governance Principles*").[1353] These principles, ranging from privacy, security, and "controllability" to "agile governance," are later reflected in the government's legislative measures on recommended algorithms, deep synthesis, and generative AI.[1354]

In September 2021, the CAC led efforts to publish the *Guiding Opinions on Strengthening Overall Governance of Internet Information Service Algorithms* ("*Guiding Opinions*").[1355] The *Guiding Opinions* outlined the plan to establish a comprehensive governance framework for algorithmic safety over approximately three years. It called for measures to establish a system of regulation and monitoring for algorithm security.[1356]

Chinese authorities, following the *Guiding Opinions*, enacted three laws, each targeting a specific technological application of AI: the *Administrative Provisions on Algorithm Recommendation for Internet Information Services*, the *Internet Information Service Deep Synthesis Management Provisions*, and the *Interim Administrative Measures for Generative AI Services*.[1357] These initiatives involve several authorities, starting with the CAC, which is the key government organ that spearheads almost all of the AI-related legislative activity. TC260, a technical work organization engaged in the formulation of information

---

[1352] Matt Sheehan, *China's AI Regulations and How They Get Made*, CARNEGIE ENDOWMENT FOR INT'L PEACE (July 10, 2023), https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

[1353] *Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI'*, DIGICHINA (June 17, 2019), https://digichina.stanford.edu/work/translation-chinese-expert-group-offers-governance-principles-for-responsible-ai/.

[1354] *Id.*; Matt Sheehan, *China's AI Regulations and How They Get Made*, CARNEGIE ENDOWMENT FOR INT'L PEACE (July 10, 2023), https://carnegieendowment.org/research/2023/07/chinas-ai-regulations-and-how-they-get-made?lang=en.

[1355] Guanyu Jiaqiang Hulianwang Xinxi Fuwu Suanfa Zonghe Zhili de Zhidao Yijian(《关于加强互联网信息服务算法综合治理的指导意见》) [Guiding Opinions on Strengthening the Comprehensive Governance of Algorithms in Internet Information Service] (promulgated by the Cyberspace Administration of China et al., effective Sept. 17, 2021), http://www.cac.gov.cn/2021-09/29/c_1634507915623047.htm.

[1356] *Translation: Guiding Opinions on Strengthening Overall Governance of Internet Information Service Algorithms*, DIGICHINA (Oct. 18, 2021), https://digichina.stanford.edu/work/translation-guiding-opinions-on-strengthening-overall-governance-of-internet-information-service-algorithms/.

[1357] Sheehan, *see supra* note 1374.

security standards,[1358] released technical standards in February 2024 for generative AI services. The standards are entitled the *Basic Safety Requirements for Generative AI Services*.[1359]

5.2.3.A. *Administrative Provisions on Algorithm Recommendation for Internet Information Services* (2021)

Following the *Guiding Opinions*, the CAC and three other ministries[1360] jointly issued the *Administrative Provisions on Algorithm Recommendation for Internet Information Services* ("*Algorithm Recommendation Provisions*"), in December 2021.[1361] The law went into effect on March 1, 2022, and introduced a regulatory framework for algorithmic recommendations used in online services in China. These online services include generative or synthetic, personalized recommendations, ranking and selection, search filter, dispatching, and decision-making.[1362]

The *Algorithm Recommendation Provisions* sets forth a comprehensive supervisory framework that details the coordinated efforts among various departments, including the national departments of cybersecurity and information, telecommunications, public security, and market regulation departments, each according to their respective responsibilities.[1363] The *Algorithm Recommendation Provisions* also introduced a system for the classification and graded management of algorithms based on factors such as their potential

---

[1358] TC260 Quanguo Wangluo Anquan Biaozhunhua Jishu Weiyuanhui (TC260 全国网络安全标准化技术委员会) [TC260 National Information Security Standardization Technical Committee], QUANGUO BIAOZHUN XINXI GONGGONG FUWU PINGTAI (全国标准信息公共服务平台) [National Public Service Platform for Standards Information], https://std.samr.gov.cn/search/orgDetailView?tcCode=TC260 (last visited May 5, 2024).

[1359] Shengchengshi Rengong Zhineng Fuwu Anquan Jiben Yaoqiu (生成式人工智能服务安全基本要求) [Basic Safety Requirements for Generative Artificial Intelligence Services] (promulgated by National Technical Committee 260 on Cybersecurity of Standardization Admin.) [hereinafter *the Basic Safety Requirements]*, NATIONAL TECHNICAL COMMITTEE 260 ON CYBERSECURITY OF STANDARDIZATION ADMIN. (Feb. 29, 2024), https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf*; translated in* Basic Safety Requirements for Generative Artificial Intelligence Services (Apr. 4, 2024), CENTER FOR SEC. & EMERGING TECH., https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai-final/.

[1360] Other institutions include the Ministry of Industry and Information Technology, the Ministry of Public Security, and the State Administration for Market Regulation.

[1361] Hulianwang Xinxi Fuwu Suanfa Tuijian Guanli Guiding (互联网信息服务算法推荐管理规定) [Administrative Provisions on Algorithm Recommendation for Internet Information Services] (promulgated by the Cyberspace Administration of China, the Ministry of Industry and Information Technology, the Ministry of Public Security, and the State Administration for Market Regulation, Dec. 31, 2021, effective Mar. 1, 2022) [hereinafter *Algorithm Recommendation Provisions*], https://www.gov.cn/zhengce/zhengceku/2022-01/04/content_5666429.htm; *translated in* Rogier Creemers, Graham Webster, Helen Toner, *Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022*, DIGICHINA (Jan 10, 2022), https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/; *Provisions on the Management of Algorithmic Recommendations in Internet Information Services* (Jan 4, 2022), CHINA LAW TRANSLATE, https://www.chinalawtranslate.com/en/algorithms/.

[1362] *Algorithm Recommendation Provisions*, art. 2 ; *China's New AI Regulations*, LATHAM & WATKINS LLP (Sept. 6, 2023), https://www.lw.com/admin/upload/SiteAttachments/Chinas-New-AI-Regulations.pdf.

[1363] *Algorithm Recommendation Provisions*, art. 3.

influence on public opinion or social mobilization, their types of content, user base size, the significance of data processed by the algorithms, and their impact on user behavior.[1364]

1) Algorithm registry

A major provision of this law requires online service providers to furnish information for a state-maintained algorithm registry.[1365] Providers of algorithmic recommendation services with "public opinion properties or having social mobilization capabilities," such as social media platforms, have to complete a form to provide various details to the government, including name, service form, domain of application, algorithm type, and algorithm self-assessment report.[1366]

2) Standards for information management

The *Algorithm Recommendation Provisions* requires online service providers to establish standards for information management. They must implement a comprehensive system covering user registration, pre-publication review of information, data security, personal information protection, and emergency handling of security incidents. Additionally, they are required to regularly audit, evaluate, and validate their algorithmic governance mechanisms, models, data, and application results. Online service providers must also conduct security assessments and maintain network records.[1367]

3) Illegal activities

The *Algorithm Recommendation Provisions* stipulates specific prohibitions. Online service providers must adhere to mainstream values and refrain from using algorithm recommendation services to engage in illegal activities or disseminate illegal information.[1368] Providers should also take measures to prevent disseminating harmful or illegal information[1369] And must not engage in various practices, including:

1. Setting up algorithmic models that induce illegal consumer behaviors[1370] or promote immoral conduct;[1371]
2. Using illegal and harmful information as user tags for pushing information[1372] or disseminating fake news;[1373]
3. Creating fake accounts to manipulate public opinion or evade regulatory oversight;[1374]

[1364] *Id.* art. 23.
[1365] *Id.* art. 24.
[1366] *Id.* art. 24.
[1367] *Id.* art. 27-28.
[1368] *Id.* art. 6.
[1369] *Id.*
[1370] *Id.* art. 21.
[1371] *Id.* art. 8.
[1372] *Id.* art. 10.
[1373] *Id.* art. 13.
[1374] *Id.* art. 14.

4. Engaging unreasonably in monopolistic practices or unfair competition;[1375]
5. Disseminating information to minors that may lead to harmful behavior that could negatively affect their physical and mental health or cause internet addiction;[1376] and
6. Engaging in price discrimination.[1377]

Online service providers must create a database for identifying illegal and harmful information and, upon its detection, take appropriate measures.[1378] Additionally, they must strengthen rules for user label management, establish mechanisms for manual intervention and user autonomy, and actively promote information that aligns with mainstream values.[1379]

4) User protection

User protection is another focal point of this regulation. Online service providers must inform users about the provision of algorithm services and disclose the principles, intended purposes, and operations of these algorithm services.[1380] Online service providers should protect users' rights to be informed and their right to choose, including the right to not be targeted based on personal characteristics and the right to turn off recommendation services entirely.[1381] Moreover, for specific protected groups, service providers must fulfill their duty to protect minors[1382] and the elderly[1383] by offering suitable services, protect workers' rights to compensation, rest, and leave,[1384] and protect consumers' rights to fair trade.[1385]

Collectively, these regulations demonstrate the government's dedication to ensuring the alignment between algorithm governance and its objectives to control online information related to China.

5.2.3.B. *Internet Information Service Deep Synthesis Management Provisions* (*Deep Synthesis Regulation*) (2022)

The Cyberspace Administration of China, the Ministry of Industry and Information Technology, and the Ministry of Public Security introduced the *Internet Information Service Deep Synthesis Management*

[1375] *Id.* art. 15.
[1376] *Id.* art. 18.
[1377] *Id.* art. 21.
[1378] *Id.* art. 9.
[1379] *Id.* art. 10.
[1380] *Id.* art. 16.
[1381] *Id.* art. 17.
[1382] *Id.* art. 18.
[1383] *Id.* art. 19.
[1384] *Id.* art. 20.
[1385] *Id.* art. 21.

Provisions ("*Deep Synthesis Regulation"*) in November 2022.[1386] It marked China's first departmental regulation aimed explicitly at "deep synthesis services." This regulation took effect on January 10, 2023.[1387]

The *Deep Synthesis Regulation* applies to activities within China that utilize "deep synthesis technologies" to provide internet information services.[1388] These are technologies used to create synthetically generated content. This includes using deep learning, virtual reality, and other synthetic algorithms to create text, images, audio, video, and other types of online information, covering a wide number of applications, such as question-and-answer dialogues, and digital simulations.[1389] The law clearly covers generative AI services, although, when the law was introduced, the major concern it sought to address was the proliferation of deepfakes.[1390]

The law imposes obligations on service providers, technical supporters providing technical support for deep synthesis services, and application distribution platforms.[1391] It also targets users who utilize deep synthesis services to create, reproduce, publish, or transmit information. The regulation grants relevant authorities the power to supervise and inspect deep synthesis services and to impose penalties.

1) Obligations for service providers and technical supporters

The *Deep Synthesis Regulation* outlines general requirements for providers and technical supporters of deep synthesis services. A few specific requirements include:

a) Safety measures

Deep synthesis service providers must implement safe and controllable technical safeguards.[1392] Service providers and technical supporters must conduct security assessments -independently or through professional institutions- if they provide models or templates capable of generating or editing biometric information or information potentially involving national security, national image, national interests, and

---

[1386] Hulianwang Xinxi Fuwu Shendu Hecheng Guanli Guiding (互联网信息服务深度合成管理规定) [Internet Information Service Deep Synthesis Management Provisions] (promulgated by the Cyberspace Administration of China, the Ministry of Industry and Information Technology, and the Ministry Public Security, Nov. 25, 2022, effective Jan. 10, 2023) [hereinafter *Deep Synthesis Regulation*], https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm; translated in *Provisions on the Administration of Deep Synthesis Internet Information Services* (Dec 11, 2022), CHINA LAW TRANSLATE, https://www.chinalawtranslate.com/en/deep-synthesis/.

[1387] Laney Zhang, *China: Provisions on Deep Synthesis Technology Enter into Effect*, LIBRARY OF CONGRESS (Apr 26, 2023), https://www.loc.gov/item/global-legal-monitor/2023-04-25/china-provisions-on-deep-synthesis-technology-enter-into-effect/.

[1388] *Deep Synthesis Regulation*, art. 2.

[1389] *Id.* art. 23; Zhang, *supra* note 1387.

[1390] This law was introduced on November 25, 2022, five days before the launch of ChatGPT. *See* Sheehan, *supra* note 1354.

[1391] Paulger, *supra* note 1326.

[1392] *Id.* art. 7.

the public interest.[1393] They must regularly audit, evaluate, and verify the mechanisms of deep synthesis algorithms.[1394]

b) Training data

Service providers and technical supporters are required to enhance the management and security of training data.[1395] If the training data include personal information, providers must comply with personal data protection regulations (*see above section 5.2.2.A.*).[1396] Furthermore, when offering editing functions for biometric data, such as faces and voices, providers and technical supporters shall prompt the users of the deep synthesis service to notify the individuals whose personal information is being edited and obtain their explicit consent.[1397]

c) User policies and management systems

Service providers must develop and publicize rules for technical supporters and users.[1398] Additionally, service providers and technical supporters must establish a user management system,[1399] including the verification of the identity of users. Service providers must carry out real identity verification of users in accordance with the law, based on mobile phone numbers, identity card numbers, Unified Social Credit Codes, or online identity authentication services. They are prohibited from offering services to users who have not undergone identity verification.[1400]

d) Content management

Service providers and technical supporters are prohibited from using deep synthesis services to create, replicate, publish, or disseminate fake news.[1401] They must promptly identify illegal and harmful information, take effective measures to address such content, and report to relevant authorities.[1402] Service providers and technical supporters must review input and output data.[1403] They must also establish, complete, and employ measures to dispel the rumors upon discovery, and make a report to the internet information departments and relevant departments. [1404]

1393 *Id.* art. 15.

1394 *Id.* art. 15.

1395 *Id.* art. 14.

1396 *Id.* art. 14.

1397 *Id.* art. 14.

1398 *Id.* art. 8.

1399 *Id.* art. 9.

1400 *Id.* ar. 9.

1401 *Id.* art. 6.

1402 *Id.* art.10-11.

1403 *Id.* art. 10.

1404 *Id.* art. 11.

e) Labeling of AI generated work

Service providers must clearly label (watermark) the generated or edited content in a reasonable location to indicate that it has been synthetically produced when such content could confuse the public.[1405] Specifically, providers must secure a conspicuous label onto the following AI-generated content that might cause confusion or mislead the public in the following situations:

1. services that simulate natural persons to generate or edit texts;
2. speech generation services, such as voice synthesis and imitations, or services that significantly change personal identification characteristics;
3. services that generate images or video of virtual persons, such as face generation, face swapping, face manipulation, and gesture manipulation;
4. generation or editing services such as immersive virtual reality, and
5. other services that have functions that generate or significantly alter information content.

In other cases, providers must include features that allow users of deep synthesis services to prominently label and alert others regarding their use of such services.[1406]

2) Specific obligations for service providers capable of influencing public opinion or mobilizing the public

Providers of algorithmic recommendation services are already required to furnish information for a state-maintained algorithm registry, as outlined by Article 24 of the *Algorithm Recommendation Provisions*.[1407] The *Deep Synthesis Regulation* expands this obligation, calling for providers of generative AI services capable of influencing public opinion or mobilizing the public to register with relevant regulators.[1408] They are also required to conduct a security assessment before launching any new products, applications, or features that could impact public opinion or mobilize the public.[1409]

The CAC determines which providers of generative AI services have "the attributes of public opinion or the capacity for social mobilization." The agency set the criteria out in the 2018 *Provisions on the Security Assessment for Internet Information Services with Characteristics of Public Opinions or Capable of Social Mobilization*. Services classified under this provision include platforms of "open forums, blogs, microblogs, chat rooms, chat groups, public accounts, short videos, webcasts, information sharing, embedded programs, and other information services," and other unidentified services that provide public opinion sharing and "have the ability to mobilize the public to engage in specific activities."[1410]

[1405] *Id.* art. 17.

[1406] *Id.* art. 17.

[1407] *Algorithm Recommendation Provisions*, art. 24.

[1408] *Deep Synthesis Regulation*, art. 19.

[1409] *Id.* art. 20.

[1410] Juyou Yulun Shuxing Huo Shehui Dongyuan Nengli De Hulianwang Xinxi Fuwu Anquan Pinggu Guiding (具有舆论属性或社会动员能力的互联网信息服务安全评估规定) [Provisions on the Security Assessment for Internet Information Services with Characteristics of Public Opinions or Capable of Social Mobilization] (promulgated by the Cyberspace Administration of China and the Ministry of Public Security, Nov. 15, 2018), ST. COUNCIL (Nov. 30, 2018), https://www.gov.cn/zhengce/zhengceku/2018-11/30/content_5457763.htm.

3) Obligations of application distribution platforms

Application distribution platforms must implement safety mechanisms, including pre-offering reviews, routine management, and emergency response.[1411] They must check deep synthesis services' security assessments and filings.[1412] They must promptly take measures to address any violations of state provisions.[1413]

4) Prohibitions for users

Deep synthesis service providers, technical supporters, and users are prohibited from using deep synthesis services for creating, reproducing, publishing, or disseminating illegal information, or engaging in illegal activities "such as those that endanger the national security and interests, harm the image of the nation, harm the societal public interest, disturb economic or social order, or harm the lawful rights and interests of others."[1414] Additionally, the *Deep Synthesis Regulation* forbids the use of technical means to delete, tamper with, or conceal watermarking.[1415]

5.2.3.C. *Interim Administrative Measures for Generative AI Services* (2023)

The CAC, in collaboration with seven other departments,[1416] unveiled the *Interim Administrative Measures for Generative AI Services* ("*Interim Measures*") in July 2023,[1417] following extensive consultations across various sectors. The *Interim Measures* is dedicated to overseeing generative AI services in China. Overall, it demonstrates a supportive stance toward the development of the generative AI industry and the innovation

---

[1411] *Deep Synthesis Regulation*, art. 13.

[1412] *Id.* art. 13.

[1413] *Id.* art. 13.

[1414] *Id.* art. 6.

[1415] *Id.* art. 18.

[1416] The Cyberspace Administration of China, the National Development and Reform Commission, the Ministry of Education, the Ministry of Science and Technology, the Ministry of Industry and Information Technology, the Ministry of Public Security, and the National Radio and Television Administration.

[1417] Shengchengshi Rengong Zhineng Fuwu Guanli Zanxing Banfa (生成式人工智能服务管理暂行办法) [*Interim Administrative Measures for Generative AI Services*] (promulgated by the Cyberspace Administration of China, the National Development and Reform Commission, the Ministry of Education, the Ministry of Science and Technology, the Ministry of Industry and Information Technology, the Ministry of Public Security, and the National Radio and Television Administration, Jul. 10, 2023, effective Aug. 15, 2023) [hereinafter *the Interim Measures*], https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm; *translated in Interim Measures for the Management of Generative Artificial Intelligence Services*, CHINA LAW TRANSLATE (July 13, 2023), https://www.chinalawtranslate.com/en/generative-ai-interim/; *China's New AI Regulations*, LATHAM&WATKINS, https://www.lw.com/admin/upload/SiteAttachments/Chinas-New-AI-Regulations.pdf (last visited May 3, 2024); Helen Toner et al., *How will China's Generative AI Regulations Shape the Future? A DigiChina Forum*, DIGICHINA (Apr 19, 2023), https://digichina.stanford.edu/work/how-will-chinas-generative-ai-regulations-shape-the-future-a-digichina-forum/.

of emerging technologies. This new law, which came into force on August 15, 2023, specifically targets generative AI technologies.

It is important to note that the draft version of the *Interim Measures* included stricter requirements which were later removed or relaxed after receiving stakeholder comments.[1418] For example, the draft version of the *Interim Measures* did not exempt research institutions, and it included technology that generates code.[1419] The draft also prohibited user profiling and required the implementation of measures to prevent the recurrence of illegal content generation within a three-month period after the content was reported.[1420] These requirements are not included in the final version of this law, suggesting, perhaps, that the regulators wanted to maintain a balance between innovation and regulation.

1) Scope of the Interim Measures

The *Interim Measures* sets forth the foundational requirements for providing "services of generating content in the form of text(s), picture(s), audio and video(s) to *the public within China* through the use of Generative AI Technologies."[1421] The expression "public within China" was not defined, but it could be concluded that the *Interim Measures* specifically targeted those entities providing generative AI services to end users residing in China.

The *Interim Measures* also provides certain key definitions in Article 22.[1422] "Generative AI technology" refers to "models and relevant technologies that have the ability to generate content such as texts, images, audio, or video." "Generative AI service providers" are defined as "organizations and individuals that use generative AI technology to provide generative AI services (including providing generative AI services through programmable interfaces and other means)." Additionally, "generative AI service users" are defined as "organizations and individuals that use generative AI services to generate content."

The *Interim Measures* does *not* apply to the deployment and use of generative AI by educational and research institutions, businesses, public cultural bodies, and related professional bodies that do not provide any generative AI services to the general public in China. The law specifically excludes activities related

---

[1418] Yan Luo et al., *China Proposes Draft Measures to Regulate Generative AI*, COVINGTON, https://www.insideprivacy.com/artificial-intelligence/china-proposes-draft-measures-to-regulate-generative-ai/ (last visited May 3, 2024).

[1419] Shengchengshi Rengong Zhineng Fuwu Guanli Banfa (Zhengqiu Yijian Gao) (生成式人工智能服务管理办法(征求意见稿)) [*Interim Administrative Measures for Generative AI Services (Draft)*] (drafted by the Cyberspace Administration of China, Apr. 11, 2023) [hereinafter *the Draft Measures*], art. 2, https://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm; *Interim Measures*, art. 2; *translated in* Translation: *Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment)* – April 2023, Seaton Huang, Helen Toner, Zac Haluza, Rogier Creemers, and Graham Webster (Apr. 12, 2023), DIGICHINA, https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/.

[1420] *Draft Measures* art. 9,11,15; *Interim Measures*, art. 9.

[1421] *Interim Measures*, art. 2 .

[1422] *Id.* art. 22.

to press and publication, film and television production, and literary and artistic creation, which are governed by other provisions.[1423]

The *Interim Measures* contains no explicit provisions that prevent Chinese businesses or consumers from using generative AI services offered by foreign providers.[1424] However, Article 20 provides that, where generative AI services provided from outside China do not comply with the *Interim Measures* and other relevant laws and regulations, the state internet information departments shall "address it."[1425] The state internet information departments include, but are not limited to, the Central Cyberspace Affairs Commission, Cyberspace Administration of China, National Administration of State Secrets Protection.

2) Obligations of generative AI service providers

The *Interim Measures* outlines the obligations of generative AI service providers in areas such as model training, content management, and user protection. Many of the key requirements in this law are close to the previously discussed regulations on algorithmic recommendation and deep synthesis technologies.[1426] However, the *Interim Measures* is more detailed.

a) General obligations

According to the *Interim Measures*, generative AI providers need to employ effective measures to increase transparency and the accuracy and reliability of AI generated content based on the characteristics of the service type.[1427] Providers should also promote the establishment of generative AI infrastructure and public training data resource platforms, collaboration and sharing of algorithm resources, increasing efficiency in the use of computing resources, and the orderly opening of public data by type and grade to expand high-quality public training data resources.[1428]

Providers and users of generative AI services should respect commercial ethics and protect commercial secrets; and, according to the *Interim Measures*, providers must not use advantages in algorithms, data, platforms, and so forth to establish monopolies or carry out unfair competition.[1429] Providers must respect the rights and interests of others, including their privacy and personal information.[1430]

[1423] *Id.* art. 2.

[1424] Yan Luo et al., *China Proposes Draft Measures to Regulate Generative AI*, COVINGTON, https://www.insideprivacy.com/artificial-intelligence/china-proposes-draft-measures-to-regulate-generative-ai/ (last visited May 3, 2024).

[1425] *Interim Measures*, art. 20.

[1426] *Interim Measures*, art. 3.

[1427] *Id.* art. 4.5 and 13.

[1428] *Id.* art. 6.

[1429] *Id.* art. 4.3.

[1430] *Id.* art. 4.4.

Throughout the AI development and deployment process, the *Interim Measures* includes explicit requirements for preventing discrimination against various protected populations.[1431] During processes such as algorithm design, the selection of training data, model generation and optimization, and the provision of services, providers of generative AI services need to take measures to prevent discrimination based on such characteristics as, race, ethnicity, faith, nationality, region, sex, age, profession, or health.[1432]

b) Training process

The *Interim Measures* requires developers of generative AI services to comply with specific requirements during the model training phase.[1433] It mandates more comprehensive training data management compared to the provisions outlined in the *Deep Synthesis Regulation*.[1434]

i) Training datasets

In addition to ensuring that the training dataset is safe and respects privacy protection, as outlined by Article 14 of the *Deep Synthesis Regulation*, the *Interim Measures* explicitly mentions requirements such as using lawful sources, not infringing on intellectual property rights, obtaining consent from personal data subjects, and increasing the quality, truth, accuracy, objectivity, and diversity of training data.[1435] Article 4.3 of the *Interim Measures* specifically outlines that provision and use of generative AI services should respect intellectual property rights.[1436] Moreover, the *Interim Measures* calls for providers to carry out pre-training, optimization training, and handling training data according to law. They must use data and foundation models that have lawful sources and, where intellectual property rights are involved, they must not infringe on intellectual property rights..[1437]

ii) Annotation of training data

In China, service providers must develop and implement clear, specific, and practical guidelines and training protocols for manually labeling data throughout the development phase. Developers must also undertake a quality assessment of their data annotation and conduct sample verification to evaluate the accuracy of the annotated content.[1438] This measure complements the safety requirements on the training dataset and can improve the precision of the training model.

c) Illegal content and activities

[1431] *Id.* art. 4.
[1432] *Id.* art. 4.2.
[1433] *Id.* art. 7.
[1434] *Id.* art. 7.
[1435] *Interim Measures*, art. 7.
[1436] *Id.* art. 4.3.
[1437] *Id.* art. 7.
[1438] *Id.* art. 8 I.

The *Interim Measures* introduces new content-related requirements for generative AI service providers. Providers are considered "content producers" of the generated content.[1439] Therefore, providers bear responsibility for content in accordance with the applicable laws.[1440]

Generative AI services may not be used to generate certain categories of content, including content that impacts national sovereignty or national security; content that advocates terrorism, extremism, or separatism; content that harms the nation's image; or content that promotes violence, obscenity, or fake information. The *Interim Measures* emphasizes the need for generative AI providers to uphold "socialist values," although the regulation does not specify what "socialist values" entail in this context.[1441] Content must respect intellectual property rights.[1442]

When dealing with users, generative AI service providers must clearly define and disclose the intended audience, context, and purposes of their services, guiding users to use generative AI legally and rationally.[1443] They must take steps to prevent any illegal activities by users, including through technical measures, such as warnings, limiting functions available to the user, and suspending user access to the service.[1444]

According to the *Interim Measures*, when generative AI service providers discover illegal content, they must take prompt measures to cease its generation and dissemination. Upon discovering illegal content, service providers must also report incidents to the relevant authorities.[1445] In addition, they must rectify the issue for the future, for example, by "optimizing training" of models to prevent the generation of illegal content.[1446]

d) User protection

Generative AI providers must clearly define and disclose their use policies.[1447] They are to provide safe, stable, and sustained services throughout to ensure users' normal usage.[1448]

When dealing with user data, generative AI providers bear responsibility as handlers of personal information and must fulfill obligations to protect personal information where personal information is involved.[1449] According to the *Interim Measures*, providers shall maintain confidentiality of users' prompts and usage records in accordance with the law. They must not collect or illegally retain personal information

[1439] *Id.* art. 9 .
[1440] *Id.* art. 9.
[1441] *Id.*
[1442] *Id.* art. 4.
[1443] *Id.* art. 10.
[1444] *Id.* art. 14.
[1445] *Id.*
[1446] *Id.* art. 14.
[1447] *Id.* art. 10.
[1448] *Id.* art. 13.
[1449] *Id.* art. 9.

from users from which a user's identity can be determined, or illegally provide users' information inputs to third parties.[1450] Providers shall lawfully and promptly accept and address requests from data subjects to have access to, reproduce, modify, supplement, or delete their personal data.[1451] Relevant entities and personnel participating in security assessments and oversight inspections of generative AI services shall keep personal privacy and personal information strictly confidential. [1452]

Generative AI service providers must establish a mechanism for receiving and handling complaints from users.[1453] They must take effective measures to prevent minors from becoming overly reliant on or addicted to generative AI services.[1454]

e) Watermarking/ tagging standard

The *Deep Synthesis* regulation already provides that, when generative AI content could confuse the public, service providers must clearly tag the generated or edited content in a reasonable location to indicate that the content has been synthetically produced.[1455] The *Interim Measures* reiterates this obligation.[1456] To implement this requirement, China's National Information Security Standardization Technical Committee ("TC260") released the final version of the *Practical Guidelines for Cybersecurity Standards – Method for Tagging Content in Generative Artificial Intelligence Services* ("*Tagging Standards*") on August 25, 2023, seeking public feedback.[1457]

The *Tagging Standards* categorizes two types of watermarking techniques to label content generated by AI. One is an "explicit watermark," to be displayed on the service interface, indicating that the content is generated by AI. The other is an "implicit watermark," embedded in the content (picture, audio, and video) in a manner that is imperceptible to humans but can be identified and extracted using technical methods. This implicit watermark must include, at a minimum, the name of the service provider. It may contain additional details, such as a unique content ID. Furthermore, when AI-generated content is downloaded as files, its metadata must encompass extra information, such as the service provider's details, the time of creation, and the unique content ID.

f) Stricter obligations for generative AI service providers capable of influencing public opinion or mobilizing the public

[1450] *Id.* art. 11.

[1451] *Id.* art. 11.

[1452] *Id.* art. 19.

[1453] *Id.* art.11, 15.

[1454] *Id.* art. 10.

[1455] *Provisions on Deep Synthesis*, art. 27.

[1456] *Interim Measures*, art. 12.

[1457] *Practical Guidelines for Cybersecurity Standards – Method for Tagging Content in Generative Artificial Intelligence Services* (网络安全标准实践指南——生成式人工智能服务内容标识方法) [promulgated by China's National Information Security Standardization Technical Committee ("TC260")] (Aug. 25, 2023), https://www.tc260.org.cn/upload/2023-08-08/1691454801460099635.pdf.

As already provided for by the *Algorithm Recommendation* and the *Deep Synthesis*, the *Interim Measures* provides that generative AI services with the capacity to affect public opinion or social mobilization must carry out security assessments and "perform formalities for the filing, modification, or canceling of filings on algorithms." [1458]

3) Enforcement of the *Interim Measures*

a) Competent authorities and possible sanctions

The *Interim Measures* outlines the supervisory responsibilities of the relevant government agencies.[1459] It authorizes these authorities to supervise and periodically inspect generative AI service providers. Providers must cooperate by providing the necessary information and technical support, by explaining "the sources, models, types, tagging rules, algorithm mechanisms, etc." for training data, and by providing necessary technical, data, and other support and assistance.[1460]

The relevant regulatory departments in charge can impose penalties on service providers in accordance with relevant laws or regulations, such as the PRC Data Security Law or the Personal Information Protection Law (PIPL). If laws or regulations are silent on a matter, the regulatory departments may issue warnings, circulate criticisms, order corrections within a specified time frame, or, in severe circumstances or cases of refusal to correct, order the suspension of generative AI services.[1461] In the case where the violation concerns public security, a public security administrative sanction may be decided. The *Interim Measures* also specifies that criminal responsibility is possible "where a crime is constituted."[1462]

If generative AI services provided from *outside* mainland China fail to comply with the requirements of laws, administrative regulations, or the *Interim Measures,* the state internet information department shall instruct relevant authorities to implement technical and other necessary measures to address the issue.[1463] Additionally, users who find that service providers have not complied with the *Interim Measures* may also file a complaint with the relevant authorities.[1464]

b) Caselaw

[1458] *Interim Measures,* art. 17.

[1459] Departments such as the Cyberspace Administration, Development and Reform, Education, Science and Technology, Industry and Information Technology, Public Security, Broadcasting and Television, and Press and Publishing shall, according to their respective duties, establish rules and enhance the management of generative AI services; *see* art. 16 of the *Interim Measures*.

[1460] *Interim Measures,* art. 19.

[1461] *Interim Measures,* art. 21.

[1462] *Id.* .

[1463] *Id.* art. 20.

[1464] *Id.* art. 18.

The Guangzhou Internet Court rendered a judgment in February 2024, applying the *Interim Measures* within the context of a copyright infringement case filed against a generative AI website.[1465] One of the issues before the Court was whether the company operating the website had met its duty of care obligations under this new law on generative AI services. The defendant company argued that it did not develop the generative AI model itself; it was using a third-party service by interfacing with the third party's application program. Thus, the defendant company argued, it should not be subject to the measures required to be taken by generative AI service providers under this new law.

The Court rejected this argument. It held that the definition of "generative AI service providers" includes "providing generative AI services through programmable interfaces and other means."[1466] The court found that the defendant company failed to comply with various obligations under this new law, including failure to implement a user complaint mechanism, under Article 15; failure to inform users about respecting intellectual property rights, under Article 4; and failure to label the AI-generated content, under Article 12.[1467] Thus, the court held, the defendant company failed to comply with its duty of care under the law and, accordingly, the court directed the defendant company to pay compensation to the plaintiff. However, the court noted in its decision that "it is inappropriate to put excessive obligations on the service providers."[1468] At the end of the decision, the court highlighted the need for a Chinese-style AI governance system that is balanced, inclusive, and compatible with both innovation and protection.[1469]

5.2.3.D. *Basic Safety Requirements for Generative AI Services* (2024)

The TC260 released the official Technical Document for *Basic Safety Requirements for Generative AI Services* ("*Basic Requirements*") on February 29, 2024.[1470] The document, released in the form of a

---

[1465] Johanna Costigan, *China Rules AI Firm Committed Copyright Infringement*, FORBES (Feb 29, 2024), https://www.forbes.com/sites/johannacostigan/2024/02/29/china-rules-ai-firm-committed-copyright-infringement/?sh=20f7cd65454d.

[1466] *Interim Measures*, art. 22.

[1467] Seagull Song & Wang Mo, *China's First Case on AIGC Output Infringement- Ultraman*, KING & WOOD MALLESONS (Feb. 28, 2024), https://www.kwm.com/global/en/insights/latest-thinking/china-s-first-case-on-aigc-output-infringement-ultraman.html.

[1468] Christine Yu & Toby Bond, *Liability of AI Service Providers for Copyright Infringement: Guangzhou Internet Court reaches world's first decision*, BIRD & BIRD (Apr. 10, 2024), https://www.twobirds.com/en/insights/2024/china/liability-of-ai-service-providers-for-copyright-infringement.

[1469] Seagull Song & Wang Mo, *China's First Case on AIGC Output Infringement- Ultraman*, KING & WOOD MALLESONS (Feb. 28, 2024), https://www.kwm.com/global/en/insights/latest-thinking/china-s-first-case-on-aigc-output-infringement-ultraman.html.

[1470] Shengchengshi Rengong Zhineng Fuwu Anquan Jiben Yaoqiu (Zhengqiu Yijian Gao) (生成式人工智能服务安全基本要求) [*Basic Safety Requirements for Generative Artificial Intelligence Services*] (promulgated by National Technical Committee 260 on Cybersecurity of Standardization Admin., Feb. 29, 2024) [hereinafter *the Basic Safety Requirements*], NATIONAL TECHNICAL COMMITTEE 260 ON CYBERSECURITY OF STANDARDIZATION ADMIN. (Feb. 29, 2024), chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf; *translated in Basic Safety Requirements for Generative Artificial Intelligence Services*, CENTER FOR SECURITY AND EMERGING TECHNOLOGY (Apr. 4, 2024), https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai-final/.

technical standard for generative AI, incorporates feedback from a previous draft document, published in October 2023.[1471] The *Basic Requirements* provides one of the most comprehensive guidelines for generative AI service providers in China to regulate training data, algorithms, and model-generated content.

A key feature of the document is the identification of 31 risks within five broad categories:

- contains content that violates the socialist core values concept;
- contains discriminatory content;
- commercial violations;
- violations of the legitimate rights and interests of others; and
- inability to meet the safety requirements of specific service types.[1472]

Certain risks, such as discrimination, underlying biases, copyrights infringements, and lack of protection for personal information, are internationally acknowledged. However, the risks associated with challenging socialist core values are specifically pertinent to the tightly regulated online environment in China.

To mitigate the 31 risks identified, the *Basic Requirements* mandates service providers to implement various assessments and measures.

1) Safety measures

A number of measures are designed to ensure the safety of generative AI models and systems. Generative AI service providers utilizing third-party foundation models for research and development must use only foundation models registered with the main oversight department.[1473] When completing filing procedures, providers must conduct safety assessments in accordance with the stipulations outlined in Article 9 of the *Basic Requirements* and submit corresponding assessment reports.[1474]

Providers should separate training and inference environments to avoid data leakage and improper access.[1475] Technical measures should be employed to enhance the model's ability to

- respond to user intent,
- align generated content with common scientific knowledge and mainstream perception,
- reduce erroneous output,
- improve the rationality of the content's format framework, and

---

[1471] Draft for Feedback: Shengchengshi Rengong Zhineng Fuwu Anquan Jiben Yaoqiu (Zhengqiu Yijian Gao) (生成式人工智能服务安全基本要求(征求意见稿))【*Basic Safety Requirements for Generative Artificial Intelligence Services (Draft for Comment)*] (promulgated by National Technical Committee 260 on Cybersecurity of Standardization Admin., Oct. 11, 2023) [hereinafter *the Draft Basic Safety Requirements*], NATIONAL TECHNICAL COMMITTEE 260 ON CYBERSECURITY OF STANDARDIZATION ADMIN. (Oct. 11, 2023), /https://www.tc260.org.cn/upload/2023-10-11/1697008495851003865.pdf; *translated in Basic Safety Requirements for Generative Artificial Intelligence Services*, CENTER FOR SECURITY AND EMERGING TECHNOLOGY (Nov. 8, 2023), https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai/.

[1472] *Basic Requirements*, art. 7.

[1473] *Basic Requirements*, art. 6.

[1474] *Id.* art. 4.

[1475] *Id.* art. 7.

- increase the percentage of valid content to ultimately provide more helpful and relevant information to users.[1476]

Providers should establish regular monitoring and evaluation processes to promptly address any identified safety issues. This includes optimizing the model through targeted instruction, fine-tuning, and reinforcement learning.[1477] Service providers must implement protective measures commensurate with the risk level for critical infrastructure and significant applications. They should formulate safety management strategies for model updates and upgrades and establish management mechanisms for conducting internal safety assessments following major updates and upgrades.[1478]

2) Training data

When collecting training data, providers must develop a management strategy for intellectual property rights and identify any significant risks of intellectual property infringement within the corpora prior to their use in training.[1479] If the corpus contains personal data or sensitive information, the provider must obtain the data subject's consent for using their personal data to train a generative AI model.[1480] When offering services through an interactive interface, providers must disclose the personal information collected and its intended uses on the website homepage, in the service agreement, and other easily accessible locations.[1481]

Generative AI providers must ensure traceability of training data.[1482] When using an open-source corpus, developers must secure an open-source license agreement or an equivalent licensing document for that source. If aggregated network addresses, data links, or similar resources are used to point to or generate additional data for the training corpus, this data should be treated the same as self-collected data.[1483] For self-collected corpora, providers must maintain detailed collection records and must not use data that others have explicitly prohibited from being collected. [1484]

Additionally, information restricted by China's cybersecurity laws, regulations, and policy documents must not be used to train generative AI models.[1485] A safety assessment is required before collecting data from a specific corpus source.[1486] For example, if the corpus contains more than 5% of "illegal or unhealthy information" as defined by the 11 categories of illegal information and 9 categories of unhealthy

[1476] *Id.* art. 6.
[1477] *Id.*
[1478] *Id.* art. 7.
[1479] *Id.* art. 5.2(b).
[1480] *Id.* art. 5.3(c).
[1481] *Basic Requirements*, art. 7(b)(2).
[1482] *Id.* art. 5.1 (c).
[1483] *Id.*
[1484] *Id.*
[1485] *Id.* art. 5.1 (a).
[1486] *Id.* art. 5.1.

information specified in the *Governance of the Online Information Content Ecosystem* ("*Governance*"),[1487] it should not be used.[1488]

Developers are required to remove illegal and harmful content from the training corpus by employing methods such as keyword filtering, classification models, and manual spot checks.[1489] The *Requirements* stipulate that the model training data must achieve a qualification rate of at least 96% through manual sampling, using a random sample of no fewer than 4,000 corpora from the total training corpus. The qualification rate represents the percentage of sampled training data that do not include the 31 identified risks.[1490] When sampling in conjunction with keyword, classification, or other methods, the training data must achieve at least a 98% qualification rate, with a random sample comprising no less than 10% of the total corpora.[1491] These assessments should, at a minimum, address scenarios involving violations of "socialist core values" or the inclusion of discriminatory content. The keyword database used to measure the qualification rate must include over 10,000 keywords. The recommended update frequency for the database is once per week.[1492]

3) Integrity of generated content

Service providers must also verify the "integrity" of the generated content with regards to the 31 identified risks.[1493] *The generated content must meet a minimum qualification standard of 90%*, regardless of the sampling method employed. The generated content test question bank should contain at least 2,000 test questions. The suggested update frequency is once per month.[1494]

[1487] This regulation, promulgated by the Cyberspace Administration of China (CAC), came into force in March 2020 and governs the industry of digital content creation and distribution, from social media to generative AI applications. *Id.*, art. 5.1(a); Wangluo Xinxi Neirong Shengtai Zhili Guiding (网络信息内容生态治理规定) [*Provisions on the Governance of the Online Information Content Ecosystem*] (promulgated by Cybersecurity Administration of China, Dec. 15, 2019, effective Mar. 1, 2020), CYBERSECURITY ADMINISTRATION OF CHINA (Dec. 20, 2019), https://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm*; translated in Provisions on the Governance of the Online Information Content Ecosystem*, CHINA LAW TRANSLATE (Dec. 21, 2019), https://www.chinalawtranslate.com/en/provisions-on-the-governance-of-the-online-information-content-ecosystem/#:~:text=%22Governance%20of%20the%20online%20information,cultivation%20and%20practice%20of%20the.

[1488] In the draft released in October 2023 seeking feedback, the National Information Security Standardization Technical Committee introduced the concept of "training data source blacklist," mandating that any source with over 5% "illegal and unhealthy information" must be blacklisted and banned from usage. Subsequently, the official requirements relax the stringent prohibition, likely due to the scarcity of high-quality data available for generative AI development.

[1489] *Basic Requirement*s, art. 5.2.

[1490] Qualification is defined as "The proportion of samples that do not include the 31 security risks listed in Appendix A of this document." *Id*., art. 9.2, 9.3, 3.4.

[1491] Qualification is defined as "The proportion of samples that do not include the 31 security risks listed in Appendix A of this document." *See Basic Requirements*, art. 3.4.

[1492] *Basic Requirements*, art. 8.1.

[1493] *Id.* art. 8.2(b).

[1494] *Id.* art. 8.2.

4) Monitoring and use of user prompts

Providers must ensure that their services are designed to avoid engaging with queries that could lead to the generation of prohibited content.[1495] Queries that are clearly biased or likely to induce the generation of "illegal or unhealthy" information must be rejected.[1496] Service providers are required to evaluate model performance regarding correct refusals to queries.[1497] They should create two databases: one for questions the models should refuse to answer and another for questions the models should always accept. The refusal database should contain at least 500 questions encompassing at least 17 risk scenarios listed by the *Basic Requirements*,[1498] including the violation of "socialist core values" or discriminatory content. The acceptance database should also contain at least 500 questions, covering topics such as the political system, beliefs, image, culture, customs, ethnicity, geography, history, national heroes and martyrs, as well as personal attributes like gender, age, occupation, and health. Models with a specific purpose may omit certain categories from the acceptance database, but these categories must still be tested in the refusal database to ensure a comprehensive evaluation of the model.[1499]

User prompts should be scrutinized using methods such as keywords and classification models.[1500] If a user inputs "illegal or unhealthy" prompts three times consecutively or accumulates five instances within a day, or if they induce the generation of "illegal or unhealthy" content, providers should take measures, such as suspending service.[1501]

Finally, data from user prompts should be used to train a model only if users have provided explicit authorization.[1502] Users should be provided with convenient methods to opt out of using their input information for training. The opt-out process should be straightforward and should require no more than four clicks from the main interface.[1503]

5) Transparency

Service providers must ensure transparency by disclosing essential information to the public and users.[1504] When providing services via an interactive interface, providers should prominently display the suitable audience, usage situations, and purposes, along with foundation model usage information. For interactive services, providers should also disclose to users the limitations, model and algorithm summaries, and

---

[1495] *Id.* art. 7(g)(2).

[1496] "Illegal and unhealthy" information is defined as a term covering 11 types of illegal information and 9 types of unhealthy information specified in the *Provisions on the Governance of the Online Information Content Ecosystem*. *Id.*

[1497] *Id.* art. 8.3.

[1498] *Id.* art. 8.3, Appendix A.1.

[1499] *Id.* art. 8.3.

[1500] *Id.* art. 7(i).

[1501] *Id.* art. 7.

[1502] *Id.* art. 5.1(a)(4).

[1503] *Id.* art. 7(c)(1).

[1504] *Id.* art. 7(b).

personal data collection and usage details in easily accessible locations, such as the website homepage or service agreement. For programmable interfaces, the above information must be included in the descriptive documentation.[1505]

6) Protection of minors

If the generative AI service is considered as not suitable for minors, then technical or administrative measures should be taken to prevent minors from having access to the service.[1506] If the service is intended for minors, guardians should be permitted to implement "anti-addiction measures." Additionally, minors should not be offered paid services that exceed their legal capacity, and they should be proactively presented with content that benefits their physical and mental health.[1507]

5.2.3.E. Toward a comprehensive AI law?

As AI technology progresses rapidly and poses general concerns beyond the confines of the specific categories of recommendation algorithms, deep synthesis, and content generation, the Chinese government seems to be considering a more comprehensive AI law. In June 2023, the State Council – the executive branch of the central government – listed a general AI law as one of its legislative projects to be included in the agenda for submission for review by the Standing Committee of the National People's Congress for 2023. In August 2023, a group of AI governance leaders from leading Chinese government-led think tanks, including the Chinese Academy of Social Sciences, released their own draft for an Artificial Intelligence Law (Model Law) 1.0.[1508] However, there has been no further official development on the introduction of a comprehensive AI law.

5.2.4 China's main initiatives on AI ethics

The country's national governance strategies for AI ethics are collectively defined by two documents. The first document, issued in September 2021, is the *Ethical Norms for New Generation Artificial Intelligence*. It was issued by the National New Generation Artificial Intelligence Governance Specialist Committee ("Specialist Committee").[1509] The second document, issued in September 2023, is the *Measures for*

[1505] *Id.*

[1506] *Id.* art. 7(a).

[1507] *Id.*.

[1508] Chinese Academy of Social Sciences Major National Conditions Research Project, Rengong Zhineng Fa (Shifan Fa) 1.0 (Zhuanjia Jianyi Gao) Qicao Shuoming He Quanwen (《人工智能法（示范法）1.0》（专家建议稿）起草说明和全文) ["Artificial Intelligence Law (Model Law) 1.0" (Expert Suggestion Draft) Drafting Instructions and Full Text], XIN ZHILI (Aug.15, 2023), https://mp.weixin.qq.com/s/XbFNnBQzoouz_q__D9Aj2g.

[1509] Xinyidai Rengong Zhineng Lunli Guifan (《新一代人工智能伦理规范》[*the Ethical Norms for New Generation Artificial Intelligence*] (promulgated by the National Professional Committee for the Governance of the New Generation of Artificial Intelligence, Sept. 25, 2021) [hereinafter the *Ethical Norms*], MINISTRY OF SCIENCE AND TECHNOLOGY, https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html; *translated in* Ethical Norms

Scientific and Technological Ethics Review (for Trial Implementation), from China's Ministry of Science and Technology and nine other government agencies. At the municipal-level, Shenzhen and Shanghai, two of China's most populous cities, have led efforts to implement relevant governance measures on AI ethics.[1510]

5.2.4.A. *Ethical Norms for New Generation Artificial Intelligence*

The Specialist Committee issued the *Ethical Norms for New Generation Artificial Intelligence* ("*Ethical Norms*") on September 26, 2021.[1511] The document seeks to incorporate ethical considerations into every developmental stage of AI technologies, applicable to individuals, legal entities, and other organizations involved in AI.[1512]

The *Ethical Norms* document establishes six foundational ethical principles: (1) advancement of human welfare, (2) promotion of fairness and justice, (3) protection of privacy and security, (4) assurance of controllability and trustworthiness, (5) strengthening accountability, and (6) improving ethical literacy.[1513]

In accordance with these principles, the Specialist Committee also introduced 18 specific ethical guidelines for the respective stages of AI's developmental life cycle: management, research and development, supply, and usage. The management guidelines emphasize agile governance, responsible demonstration, proper exercise of power, risk prevention, and promotion of inclusiveness and openness. The research and development guidelines call for self-discipline, data quality improvement, safety and transparency enhancement, and avoiding bias and discrimination. The supply guidelines advocate for respecting market rules, quality control, protecting users' rights and interests, and emergency protection. Lastly, the *usage norms* encourage goodwill, caution against misuse and abuse, outlaw illegal and malicious use, and emphasize the importance of timely feedback and improving usage skills.[1514]

Although the *Ethical Norms* document defines what these ethical guidelines entail, it has neither specified relevant enforcement mechanisms nor addressed the ramifications in the face of violations.[1515]

5.2.4.B. The *Measures for Scientific and Technological Ethics Review (Trial Measures)*

for New Generation Artificial Intelligence Released (Oct. 21, 2021), CSET, https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/.

[1510] Ashyana-Jasmine Kachra, *Making Sense of China's AI Regulations*, HOLISTIC AI (Feb. 12, 2024), https://www.holisticai.com/blog/china-ai-regulation.

[1511] *Ethical Norms*, *see supra* note 1509.

[1512] *Ethical Norms*, art.1-2.

[1513] *Id.*, art. 3.

[1514] *Ethical Norms*, art.5-9 (for management); art.10-13 (for R&D); art.14-17 (for supply); art.18-22 (for use).

[1515] *Id.*

The Ministry of Science and Technology ("MOST") and nine other government agencies[1516] introduced the *Measures for Scientific and Technological Ethics Review (for Trial Implementation)* ("*Trial Measures*") on September 7, 2023.[1517] The *Trial Measures* mandates the ethical review of scientific and technological activities involving (1) humans as research participants and the use of human biological samples and personal information data, (2) experimental animals, (3) ethical challenges in terms of life and health, ecological environment, public order, sustainable development, and (4) other scientific and technological activities that must undergo an ethical review in accordance with laws, administrative regulations, and relevant national regulations.[1518]

The *Trial Measures* specifies the entities responsible for conducting ethical reviews and outline the specific procedures for such reviews. The review entities include higher education institutions, scientific research institutions, healthcare institutions, and enterprises, especially those engaged in life sciences, medicine, AI, and other scientific activities that touch upon ethically sensitive areas.[1519] The entities are responsible for conducting ethical reviews and, especially for institutions in the field of AI, they are required to establish internal scientific ethics committees.[1520] These committees are tasked with a broad range of responsibilities, from developing management systems and working standards to conducting ethical reviews and providing guidance on ethical risk assessments.[1521]

The procedures for ethical review are categorized into four types: "general procedures," "simplified procedures," "expert review procedures," and "emergency procedures," based on evaluation of the ethical risks associated with the technological activities under scrutiny.[1522] However, while the *Interim Measures* included "graded management by categories of generative AI services" as a guiding principle for regulating AI, no definitions of such grades or categories are codified in the *Trial Measures* regulation.

MOST supervises ethical reviews nationwide. This includes maintaining lists of noncompliant entities, individuals, and ethics committees. Additionally, the *Trial Measures* introduces a national information registration platform dedicated to managing ethical reviews. This platform specifies the registration

---

[1516] The Ministry of Science and Technology, the Ministry of Education, the Ministry of Industry and Information Technology, the Ministry of Agriculture and Rural Affairs, the National Health Commission, Chinese Academy of Sciences, Chinese Academy of Social Sciences, Chinese Academy of Engineering, China Society and Technology Association, and Chinese Military Commission of Science and Technology.

[1517] Guanyu Yinfa "Keji Lunli Guanli Banfa (Shixing)" de Tongzhi (关于印发《科技伦理审查办法（试行）》的通知) [*Notice of Issuing the Trial Measures for Scientific and Technological Ethics Review*] (promulgated by the Ministry of Science and Technology, the Ministry of Education, the Ministry of Industry and Information Technology, the Ministry of Agriculture and Rural Affairs, the National Health Commission, Chinese Academy of Sciences, Chinese Academy of Social Sciences, Chinese Academy of Engineering, China Society and Technology Association, and Chinese Military Commission of Science and Technology, effective Sep. 7, 2023) [hereinafter *Trial Measures*], https://www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm; *translated in* https://lawinfochina.com/display.aspx?id=42015&lib=law&SearchKeyword=&SearchCKeyword=.

[1518] *Trial Measures*, art. 2.

[1519] *Id.* art. 4.

[1520] *Id.* art. 5.

[1521] *Id.* art. 5.

[1522] *Id.* art. 3.

requirements for entities, catalogs scientific and technological activities under review, and mandates the submission of annual committee reports.[1523]

Institutions or ethical review committees that fail to comply can face administrative, civil, or criminal consequences based on the severity of the violations committed. These behaviors include forging ethical review approvals, obtaining ethical review approvals through deceit or fake evidence, interference with the ethical review process, and conducting scientific and technological activities without a valid ethical review approval.

Therefore, the *Trial Measures* is pivotal in delineating the ethical governance guidelines for scientific and technological activities, with particular implications for AI. By requiring entities to establish internal ethics review committees, this comprehensive framework institutionalizes a rigorous ethical oversight mechanism. By systematizing ethical reviews and instituting a registration platform to monitor compliance, China signals it is dedicated to enforcing its ethical standards in science and technology, safeguarding societal welfare against the backdrop of rapid technological advancement.

5.2.4.C Municipal-level AI ethics committee

Some provinces and municipalities in China have taken proactive measures to define their own rules to promote the safe development of AI. Since late 2022, the municipal and provincial governments of Shenzhen, Shanghai, Guangdong, Jiangsu, and Zhejiang have each passed regulations. These regulations define the municipal government's role in promoting the development of the AI industry and setting the principles for municipal-level AI governance. Among these policy initiatives, the *Regulations for the Promotion of the Artificial Intelligence Industry in Shenzhen Special Economic Zone* ("*Shenzhen Regulations"*) and the *Regulations for the Promotion of the Development of the Artificial Intelligence Industry in Shanghai Municipality* set prominent examples for local AI governance through the establishment of the municipal-level AI ethics committees. The *Shenzhen Regulations* is a useful case study to show how local governments are employing specific measures to regulate the ethical risks of AI technologies.

The *Shenzhen Regulations* outlines six major responsibilities for the planned AI ethics committee, including establishing the ethical norms, the foundational management system for such norms, and evaluating the implementation of these norms.[1524] In addition, the evaluation is to follow a differentiated mechanism based on categorization and grading of the AI technology's relevant risks.[1525] For example, the framework

[1523] *Id.* art. 40, 42, 43, 44, 45.

[1524] Standing Comm. Shenzhen Mun. People's Cong., *Article 65: Regulations for the Promotion of the Artificial Intelligence Industry in Shenzhen Special Economic Zone*, CSET (Dec. 15, 2022), https://cset.georgetown.edu/publication/regulations-for-the-promotion-of-the-artificial-intelligence-industry-in-shenzhen-special-economic-zone/.

[1525] Standing Comm. Shenzhen Mun. People's Cong., *Article 66: Regulations for the Promotion of the Artificial Intelligence Industry in Shenzhen Special Economic Zone*, CSET (Dec. 15, 2022), https://cset.georgetown.edu/publication/regulations-for-the-promotion-of-the-artificial-intelligence-industry-in-shenzhen-special-economic-zone/.

requires high-risk applications to go through "ex-ante assessment and risk warning," while mid-to-low-risk applications go through "ex-ante disclosure and ex post-facto tracking."[1526] Although the *Shenzhen Regulations* document does not clarify the categorization and grading of relevant risks, it enumerates several socio-economic factors for both individuals and organizations to use as benchmarks in evaluating AI's impact. These factors include behaviors, income changes, social psychology of individuals and organizations, and comprehensive impacts on employment structure, social equity, and other aspects.

KEY TAKEAWAYS

- In its 2017 *New Generation AI Development Plan*, China announced its objective to establish itself as a world leader in artificial intelligence by 2030. China's ambitions have since expanded. In October 2023, President Xi Jinping introduced the *Global AI Governance Initiative of China,* marking a strategic effort to influence the global governance of artificial intelligence. This initiative underscores China's commitment to fostering international cooperation to prevent the misuse of AI. It also emphasizes the importance of ensuring that developing countries have a significant voice in global AI governance, allowing all nations, regardless of economic status or political systems, to participate in AI development and oversight.

- The large-scale deployment of generative AI tools has led Chinese authorities to adopt increasingly precise and stringent regulatory frameworks. The Chinese government has successively enacted three major laws: the *Administrative Provisions on Algorithm Recommendation for Internet Information Services* (2021), the *Internet Information Service Deep Synthesis Management Provisions* (2022), and the *Interim Administrative Measures for Generative AI Services* (2023). These legislative measures complement existing regulations on data protection and copyright law. They build upon one another, with modifications to accommodate the latest iterations of AI technologies. Recently, they were complemented by technical standards for generative AI through the *Basic Safety Requirements for Generative AI Services* (2024). Additionally, this comprehensive legal framework is supplemented by the government's publication of ethical principles, such as the *Ethical Norms for New Generation Artificial Intelligence* (2021).

- From the perspective of data protection, it is important to note that Chinese law allows data controllers to process publicly available personal data to a reasonable extent without the data subject's consent, provided the data were disclosed by the data subject themselves or were otherwise legally disclosed. In other situations, the data subject's consent is generally required to use their personal data for training an AI model.

- From the perspective of copyright law, the Chinese legal system currently offers innovative and original judicial solutions, compared to other legal systems. While Chinese courts do not grant

[1526] Standing Comm. Shenzhen Mun. People's Cong., *Article 67: Regulations for the Promotion of the Artificial Intelligence Industry in Shenzhen Special Economic Zone*, CSET (Dec. 15, 2022), https://cset.georgetown.edu/publication/regulations-for-the-promotion-of-the-artificial-intelligence-industry-in-shenzhen-special-economic-zone/.

intellectual property rights to purely AI-generated works that lack originality, they now recognize copyrights for users who create original works with the assistance of generative AI. A judicial decision justified this protection by noting that users contribute to the creation process through their "prompts." Chinese courts also sanction generative AI developers when their tools produce outputs that closely resemble or duplicate copyrighted content. Additionally, a recent Chinese decision recognized the protection of an individual's voice under personality rights.

- The various laws adopted since 2021 focus on algorithmic recommendations used in online services, technologies for creating synthetically generated content ("deep synthesis"), and generative AI services. These laws primarily reflect the Chinese government's chief concern to prevent illegal activities under Chinese law, such as the generation and dissemination of "fake news" and "illegal or unhealthy information." This objective justifies a significant number of restrictive measures targeting both service providers and users. Users must be identified, and their online activities closely monitored. Generative AI tools must be trained and configured to minimize the generation of prohibited content. Training data and generated outputs are subject to content monitoring, particularly to prevent "illegal or unhealthy information" from being included in the training data. Service providers must promptly address illegal content, halt its generation and transmission, and report incidents to the relevant authorities. They must also address the issue for the future, for example, by "optimizing training" of AI models to prevent the generation of illegal content.

- The provisions adopted by the Chinese government also reflect several emerging priorities. Some measures aim to ensure cybersecurity and the efficient operation of generative AI tools. Others focus on safeguarding users by enforcing personal data protection, transparency, and non-discrimination. Notably, Chinese legislators are among the few that address the risk of addiction to digital tools, especially concerning minors. Furthermore, the Chinese regulations mandate compulsory labeling of all synthetically generated or edited content. When such content might confuse the public, providers must label it conspicuously. However, when the content does not pose a risk of confusion, providers are required only to include features that allow users to prominently label it. Overall, many of the risks and challenges examined in Chapter 3 are addressed by Chinese regulations (*see Figure 45 below*). However, it is noteworthy that environmental concerns are absent from the provisions considered.

- From a methodological perspective, the Chinese legislators appear to leave little room for self-regulation or co-regulation strategies. The Chinese approach is mostly state-led, with rules formulated and specified by the national government and its agencies.

- Finally, similar to the EU AI Act, the Chinese regulatory framework targets both domestic and international service providers whose products are accessible to domestic users. However, unlike the EU AI Act, their extraterritorial reach is likely limited, as many leading global AI companies, such as OpenAI, do not operate in mainland China. Additionally, Chinese regulators primarily focus on the domestic market and do not specifically address the situation of Chinese service providers offering AI technologies to overseas users.

FIGURE 45 How the Chinese legal framework addresses identified risks

Possible risks and challenges of generative AI	Main provision of Chinese legal frameworks governing generative AI services
Technical vulnerabilities (section 3.1.1)	• Deep Synthesis service providers must implement technical safeguards and conduct security assessments (Article 7, *Deep Synthesis*) and regularly audit deep synthesis algorithms. (Article 15, *Deep Synthesis*) (*section 5.2.3.B*) • When manual tagging is conducted during research and development of generative AI technology, providers of generative AI services must formulate tagging rules, assess the quality of the tagging, and spot check the accuracy of the tagging. (Article 8, *Interim Measures*) (*section 5.2.3.C*) • Providers of generative AI services should establish regular monitoring and evaluation measures. (Article 6, *Basic Requirements*) (*section 5.2.3.D*) • Providers of generative AI services must ensure standards for the source, quality, and safety of the training corpus are met. (Article 5.1, *Basic Requirements*) (*section 5.2.3.D*) • Providers of generative AI services related to critical infrastructure and important applications must take risk-appropriate protective measures. (Article 7, *Basic Requirements*) (*section 5.2.3.D*) • Providers of generative AI services must conduct safety assessments and submit assessment results when performing filing procedures, and must establish safety management strategy for model updates and upgrades, and conduct assessment after those updates and upgrades. (Article 4, 7, and 9, *Basic Requirements*) (*section 5.2.3.D*)
Factually incorrect content (section 3.1.2.)	• Providers of generative AI tools must formulate and perform annotation rules compliant with the Measures, and spot check to evaluate the validity of annotation content. (Article 8, *Interim Measures*) (*section 5.2.3.C*) • Providers of generative AI services must perform a mandatory review of the training data (Article 5.3, *Basic Requirements*). (*section 5.2.3.D*)
Opacity (section 3.1.3)	• Providers of online services must enhance the transparency and understandability of search, ranking, selection, push notification, and display algorithms. (Article 12, *Algorithm Recommendation*) (*section 5.2.3.A*) • Generative AI service providers must disclose information and provide assistance to relevant authorities conducting inspections of generative AI services. (Article 19, *Interim Measures*) (*section 5.2.3.C*) • Generative AI service providers must disclose essential information to the public and users. (Article 7(b), *Basic Requirements*) (*section 5.2.3.D*)

Misuse and abuse (section 3.2.1)	• Deep synthesis providers and technical supporters must implement a user management system to verify user identities. (Article 9, *Deep Synthesis*) (*section 5.2.3.B*) • Deep synthesis providers and technical supporters must review input and output data. (Article 10, *Deep Synthesis*) (*section 5.2.3.B*) • Users of deep synthesis services are prohibited from creating, reproducing, publishing, or disseminating illegal information, or engaging in illegal activities. (Article 6, *Deep Synthesis*) (*section 5.2.3.B*) • Generative AI service providers bear responsibility as the producers of online information content. (Article 9, *Interim Measures*) (*section 5.2.3.C*) • Generative AI services should not generate content related to certain categories, such as terrorism, promotion of violence, obscenity, fake information, and copyright infringement, etc. (Article 4, *Interim Measures*) (*section 5.2.3.C*) • Generative AI service providers must guide users to use generative AI legally and rationally (Article 10, *Interim Measures*) and prevent any illegal activities by users. This, includes through technical measures, such as warnings; limiting functions available to the user; and suspending user access to the service. (Article 14, *Interim Measures*) (*section 5.2.3.C*) • When generative AI service providers discover illegal content, they must take prompt measures to cease its generation and dissemination. This includes stopping the generation and transmission of illegal content, removing the content, and optimizing the model to make corrections. Providers must also report incidents to the relevant authorities. (Article 14, *Interim Measures*) (*section 5.2.3.C*) • In case of generation of illegal content, generative AI providers must rectify the issue for the future**,** for example by "optimizing training" of AI models. (Article 14, *Interim Measures*) (*section 5.2.3.C*)
Misinformation and disinformation (section 3.2.2)	• Deep synthesis service providers and users are prohibited from using deep synthesis services to create, replicate, publish, or disseminate fake news. (Article 6, *Deep Synthesis*) (*section 5.2.3.B*) • Deep synthesis service providers and technical supporters must promptly identify illegal, negative, and false information; take effective measures to address such content; and report to relevant authorities. (Article 10 and 11, *Deep Synthesis*) (*section 5.2.3.B*) • Deep synthesis and generative AI service providers must conspicuously label (watermark) the generated or edited content in a reasonable location to indicate that it has been synthetically produced, when such content could confuse the public. (Article 17, *Deep Synthesis*; Article 12, *Interim Measures*) (*section 5.2.3.C*) • When the generated or edited content *cannot* confuse the public, deep synthesis service providers must include features that

	allow users to prominently label and alert others regarding their use of such services. (Article 17, *Deep Synthesis*) (*section 5.2.3.B*) ● Generative AI providers of services capable of influencing public opinion or mobilizing the public must register with relevant regulators. (Article 19, *Interim Measures*) (*section 5.2.3.C*) ● Generative AI providers shall bear responsibility as the producers of online information content in accordance with law and are to fulfill the online information security obligations. (Article 9, *Interim Measures*) (*section 5.2.3.C*) ● Generative AI providers must conduct an assessment to limit "illegal or unhealthy information" included in the training data before training. (Article 5.1, *Basic Requirements*) Providers must assess the safety of the generated content based on the requirements. (Article 8.2 and 9.3, *Basic Requirements*) (*section 5.2.3.D*) ● Generative AI tools must reject queries that contain the dissemination of false and harmful information. (Article 8.3 and Appendix 1(g), *Basic Safety Requirements*). User input must be monitored to detect illegal or unhealthy information. (Article 7(g), *Basic Requirements*) (*section 5.2.3.D*)
Bias and discrimination (section 3.2.3)	● Providers of generative AI services must take measures to prevent the creation of discrimination, such as by race, ethnicity, faith, nationality, region, sex, age, profession, or health. (Article 4.2, *Interim Measures*) (*section 5.2.3.C*) ● Prompts likely to generate discriminatory content must be rejected. (Article 8.3, *Basic Requirements*) (*section 5.2.3.D*)
Influence, overreliance, and dependence (section 3.2.4)	● Providers of algorithm services must not set up algorithmic models that violate laws and regulations or ethics and morals, such as by leading users to addiction or excessive consumption. (Article 8, *Algorithm Recommendation*) (*section 5.2.3.A*) ● Providers of algorithm services must not use algorithmic recommendation services to lead minors to online addiction. (Article 18, *Algorithmic Recommendation*) (*section 5.2.3.A*) ● Generative AI providers must take effective measures to prevent minors from becoming overly reliant on or addicted to generative AI services. (Article 10, *Interim Measures*) (*section 5.2.3.C*) ● If the service is suitable for minors, guardians shall be allowed to set up anti-addiction measures for minors. (Article 7(a), *Basic Requirements*) (*section 5.2.3.D*)
Privacy and data protection (section 3.3.1)	● Deep synthesis providers must comply with data protection law. (Article 14, *Deep Synthesis*) (*section 5.2.3.B*) ● Providers must respect the rights and interests of others, including their privacy and personal information. (Article 4.4, *Interim Measures*) (*section 5.2.3.C*) ● Generative AI providers are personal information handlers and must protect personal information. (Article 9, *Interim Measures*) (*section 5.2.3.C*)

	• Generative AI providers must not collect unnecessary personal information from users, illegally retain users' information inputs from which a user's identity can be determined, or illegally provide users' information inputs to third parties. (Article 11, *Interim Measures*) (*section 5.2.3.C*) • Generative AI providers must establish a mechanism for receiving and handling complaints from users. They should promptly address individuals' requests to access, copy, correct, supplement, or delete personal information (Article 11 and 15 *Interim Measures*) (*section 5.2.3.C*) • Generative AI providers must obtain individual consent and comply with laws before using personal or sensitive information for training. (Article 5.2 (c), *Basic Requirements*) (*section 5.2.3.D*) • Generative AI providers must disclose the personal information collected and its intended uses. (Article 7(b)(2), *Basic Requirements*) (*section 5.2.3.D*) • Data from user input should be used only to train a model with user authorized records.(Article 5.1, *Basic Requirements*) (*section 5.2.3.D*)
Copyrights (section 3.3.2)	• Generative AI service providers must respect intellectual property rights (Article 4.3, *Interim Measures*) and train their models without infringement on those rights. (Article 7, *Interim Measures*) (*section 5.2.3.C*) • Generative AI service providers must establish an intellectual property rights management strategy and identify any significant intellectual property infringement risks within the corpora. (Article 5.2, *Basic Requirements*) (*section 5.2.3.D*)
Concentration of market power (section 3.4.1)	• The provision and use of generative AI services should respect commercial ethics and must not be used for monopolies or to carry out unfair competition. (Article 4.3 *Interim Measures*) (*section 5.2.3.C*) • Generative AI providers' commercial violations include the use of algorithms, data, platforms, etc. to engage in monopolistic or unfair competition behaviors. (Appendix 3(d), *Basic Requirements*) (*section 5.2.3.D*)

5.3. The United States

Unlike the European Union or China, the United States has not implemented a comprehensive federal framework to govern artificial intelligence through mandatory rules. Instead, the federal government has primarily engaged in dialogue with major AI companies to secure commitments and encourage adherence to voluntary standards set by federal agencies. Meanwhile, numerous bills have been introduced in Congress, and several state laws have been enacted which directly address AI.

This section will begin by examining the legal provisions currently applicable to the development and use of generative AI, encompassing areas such data protection frameworks, intellectual property, and civil liability. It will then analyze the federal government's policy, highlighted by President Biden's Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence adopted on October 30, 2023.[1527] Lastly, it will explore the legislative texts introduced or enacted across various states.

5.3.1 Existing legal frameworks

At present, no US federal law, either partial or comprehensive, directly regulates artificial intelligence. In terms of data privacy and personal data protection, only a few data protection and privacy laws enacted by individual states may apply to developers of generative AI. US copyright and patent law govern the training of generative AI models and their outputs. Additionally, theories of general liability involving the First Amendment of the U.S. Constitution and Section 230 of the Communications Decency Act may determine legal responsibility for harmful or offensive outputs produced by generative AI models.

5.3.1.A. Data protection issues in the US

Unlike many other countries, the United States does not have a national comprehensive data privacy or data protection law. Several federal laws offer partial privacy protections for certain categories of data or sectors, such as the Children's Online Privacy Protection Act (COPPA) for children's data, and the Health Insurance Portability and Accountability Act (HIPAA) for healthcare data.[1528] More protection is offered at the state level to citizens of certain states; as of early 2024, nearly one-third of the 50 states have passed general

[1527] Executive Office of the President [Joseph Biden]. Executive Order #14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Fed. Reg. 75191, 75191-75226 (Nov. 1, 2023), https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[1528] Jennifer King & Caroline Meinhardt, *Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World*, STANFORD HAI (Feb. 2024) at 2, https://hai.stanford.edu/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world.

consumer data privacy laws.[1529] These state laws are broadly inspired by the European Union's GDPR, but they also differ from GDPR in ways that are relevant to how generative AI companies collect and use data.

1) Scope of data covered and the "Publicly Available Data" exemption

The privacy laws enacted by various US states do not cover all types of data. They are typically cabined by with "personal data" or "personal information." To abstract away from subtle distinctions between these terms across states, this report will refer to both by using a predecessor umbrella term, "personally identifiable information" (PII). For purposes of this report, PII refers to "information that is linked or reasonably linkable to an identified or identifiable natural person."[1530] (see section 3.3.1.A).

One relevant limitation of the definitions of PII in state privacy laws in the United States pertains to "publicly available" information. This includes information that a business has a reasonable basis to believe has already been lawfully made available to the public by the identifiable person or through widely distributed media. "Publicly available" information generally does not qualify as protected PII.[1531] This carve-out provides wide latitude for web scraping of information already on the internet. There are bills regarding data privacy pending before the U.S. Congress, such as the American Data Privacy and Protection Act (ADPPA), that contain a similar exemption.[1532] Consequently, personal information scraped from publicly accessible internet sources are unlikely to be protected under US state or federal privacy laws. This means that individuals whose data are scraped may not receive the standard notice and consent protections typically required when their data are collected from the web.

2) Notice and consent requirements generally

Though notice and consent requirements will generally not be an obstacle for training AI models on publicly available data, as a general matter, state data privacy laws impose a notice requirement on data processors who collect and use data. Businesses are obliged to both provide notice of the types of data they collect from customers and procure consent to that data collection and use.[1533] An illustrative example of this comes from the Virginia Consumer Data Protection Act (VCDPA), which took effect in January 2023. It requires companies that process personal data to provide consumers with a reasonably accessible, clear, and meaningful privacy notice. The notice must include:

[1529] For an up-to-date count, *see* Andrew Folks, *US State Privacy Legislation Tracker*, INT'L ASSOCIATION OF PRIVACY PROFESSIONALS [hereinafter IAPP] (last updated Jul. 22, 2024), https://iapp.org/resources/article/us-state-privacy-legislation-tracker/. Many of these laws are not yet effective and will come into force in the future.

[1530] The quoted language is from the VCDPA, but it is broadly representative of how PII is defined in other US state laws. Virginia Consumer Data Protection Act (VCDPA), Va. Code Ann. § 59.1-575 (West 2023). While "personally identifiable information" (PII) and personal data both refer to information that can identify an individual, PII is a narrower concept primarily used in the United States, whereas "personal data" is a broader concept used in the EU and other regions with comprehensive data protection regulations.

[1531] *See, e.g.*, Cal. Civ. Code § 1798.140 (v)(2) (West 2023); Va. Civ. Code. § 59.1-575 (2023).

[1532] American Data Privacy and Protection Act, H.R. 8152, 117th Cong. (2022).

[1533] Jennifer King & Caroline Meinhardt, *Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World*, STANFORD HAI (Feb. 2024) at 13, 33 (focusing on California's CCPA), https://hai.stanford.edu/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world.

- the categories of personally identifiable information being processed by the data controller,
- the purpose(s) of the processing,
- instructions on how consumers can exercise their rights to privacy (see below),
- the categories of PII shared with third parties, and
- what categories of third parties that PII is to be shared.[1534]

Other state privacy laws in the US require companies and AI users to give notice to consumers *and* obtain affirmative consent from them for use of their PII. The consent requirement is generally required when the PII in question is categorized as "sensitive data" or is being used for a purpose different from what was initially disclosed to consumers when the data were provided. The Colorado Privacy Act, for instance, lists the following categories as "sensitive data": race/ethnicity, religion, mental/physical health condition, sex life or sexual orientation, citizenship status, genetic or biometric data used to identify an individual, and data from a known child.[1535]

Notice and consent requirements may also come into play when training data come from information collected directly from consumers by AI companies themselves or their business partners. In these situations, companies may be required to give consumers notice and, in some situations, obtain their consent (e.g., if the company wishes to use already-collected data for the new purpose of AI training). In general, this will not present particularly novel privacy law compliance challenges, as many companies have well-established procedures by which they notify and obtain consent from existing customers to new legal terms/policies. However, when this repurposing of data involves *selling* data to third parties (e.g., one AI company buys another company's data to use in the AI company's training), compliance could be complicated by the need to respect individual customer's exercise of "do not sell my data" rights contained in many state privacy laws.

3) The rights of individuals to opt out or correct data

Following in the footsteps of the European Union's GDPR, US state privacy laws grant certain "data subject rights" to individuals (data subjects) with regard to the data that companies have collected about them. Data subject rights vary somewhat among different state laws, but examples include rights to obtain a copy of one's data, correct errors in the data, and request deletion of one's data from a company's database.

For information that is attached to accounts that the consumer has with the generative AI company itself (e.g., email, billing details), the request to opt out or correct information is quite straightforward for both user and company. Similarly, it should be relatively straightforward for companies to fulfill consumers' rights, included in many US state privacy laws, to opt out of advertising targeted at them based on their web browsing activities (so-called "behavioral advertising"). However, matters become more complicated for other types of PII that generative AI companies may hold (even without knowing it), such as PII that

[1534] § 59.1-578 (C).

[1535] Colo. Rev. Stat. § 6-1-1303 (2024). Other data privacy laws specifically protect biometric data, such as the Illinois Biometric Information Privacy Act. This law formed the basis of a class action settlement against Clearview AI for its use of facial recognition software. Chris Burt, *Clearview AI reaches preliminary deal to settle biometric data privacy lawsuit*, BIOMETRIC UPDATE (Dec. 4, 2023), https://www.biometricupdate.com/202312/clearview-ai-reaches-preliminary-deal-to-settle-biometric-data-privacy-lawsuit.

has been included in scraped datasets used for pre-training AI models or proprietary datasets used for fine-tuning.

Generative AI raises novel questions about the scope of rights for data subjects, such as whether or to what extent rights—such as the right to correct or delete personal data—would extend to information about a person that is inferred by an AI model. For example, should an individual be able to demand, under their right to deletion, that the AI model be blocked from including any information about them in its responses to AI users' queries? These are the sorts of difficult questions that companies and regulators will have to grapple with as they determine how data subject rights that pre-date generative AI should apply to this new technology.

4) Profiling

As companies build generative AI into more services, much of their functionality—for instance as personal assistants or advisors– could depend on the degree to which they are able to gain access to or infer information about individual users that can then be used to tailor helpful, personalized responses.[1536] This may bring them into contact with US state privacy laws that have provisions related to processing data for "profiling" of individuals. Profiling, as defined in representative language from the Colorado Privacy Act, is "any form of automated processing of personal data to evaluate, analyze, or predict personal aspects concerning an identified or identifiable individual's economic situation, health, personal preferences, interests, reliability, behavior, location, or movements."[1537]

Profiling is particularly important because many state privacy laws give individuals the right to opt out of profiling used in furtherance of automated decision-making that has legal or similarly significant consequences. As generative AI tools are incorporated into systems that affect important areas of people's lives—such as healthcare, finances, and employment—some of these areas, in combination with profiling, could prompt many individuals to exercise their opt-out rights.

Future enforcement actions and court cases are likely to provide guidance on how statutory terms like "profiling" and "legal or similarly significant effects" should be interpreted. Additionally, as states operationalize their privacy statutes through rulemaking, differences in the nature and degree of human involvement may be legally important. For example, the rules for the Colorado Privacy Act distinguish among three types of automated processing: Solely Automated Processing, Human Reviewed Automated Processing, and Human Involved Automated Processing.[1538] Companies using automated decision-making systems in the first two categories must honor consumer opt-out requests (when the decision-making produces legal or other significant effects). They are not required to do so for the third category, human-*involved* automated processing, where the company can more suitably explain its reasoning.

[1536] *See, e.g.*, Kyle Bradshaw, *Google Bard (now Gemini) Readies 'Memory' to Adapt to Important Details About You*, 9TO5GOOGLE (Sept. 29, 2023), https://9to5google.com/2023/09/29/google-Bard (now Gemini)-memory/.
[1537] Colo. Rev. Stat. § 6-1-1303 (20).
[1538] Colo. Code Regs. § 904-3 (2015), https://coag.gov/app/uploads/2022/10/CPA_Final-Draft-Rules-9.29.22.pdf.

5) Automated decision-making

A further challenge with automated decision-making is transparency or explainability. The Colorado Privacy Act's Rule 9.03 requires, among other things, that companies using profiling in impactful settings (e.g., housing, employment, insurance) provide consumers with a clear explanation in the privacy notice of how profiling is used. This must include a plain language explanation of the logic used in the profiling process and why profiling is relevant to a decision the company must make.[1539] Such disclosure and explanation requirements could be difficult to satisfy for generative AI tools, such as an AI model that evaluates resumes, cover letters, and other written materials to screen out job applicants. The workings of advanced AI models are notoriously opaque even to the experts developing and deploying them.[1540]

5.3.1.B. Intellectual property: copyright and patentability issues

Other emerging legal issues are novel and may be determined only through ad hoc judicial decisions over the next several years.[1541] Intellectual property issues have so far been the tip of the spear for AI in the US, particularly copyright.

1) US copyright issues related to inputs: The fair use debate

Over a dozen lawsuits were filed during the past 18 months by authors, artists, and media companies, alleging that generative AI companies had violated US copyright law by training AI models on copyrighted works without permission from or compensation to the creators of those materials (*see section 3.3.2*).[1542] No federal court has yet issued a final ruling on the merits of the lawsuits, though at least one federal court has hinted that such direct copyright infringement claims have prima facie merit.[1543] This stands to reason. As the training of models does involve scraping and using copyrighted works (by literally making intermediate copies of them), the fate of these direct copyright infringement claims is likely to turn on whether such use is considered a "fair use."[1544]

[1539] *Id.*, at § 9.03.

[1540] *See, e.g.*, Eva Eigner & Thorsten Handler, *Determinants of LLM-assisted Decision-Making*, https://arxiv.org/html/2402.17385v1; Ksenia Se, *Explainable AI And Prompting a Black Box in the Era of Gen AI*, HACKERNOON (Mar. 5, 2024), https://hackernoon.com/explainable-ai-and-prompting-a-black-box-in-the-era-of-gen-ai#.

[1541] *See, e.g.*, Christopher Mims, *The AI Industry is Heading Towards a Legal Iceberg*, WALL ST. J. (Mar. 29, 2024), https://www.wsj.com/tech/ai/the-ai-industry-is-steaming-toward-a-legal-iceberg-5d9a6ac1?mod=tech_lead_pos5.

[1542] Rachel Kim, *Copyright Alliance, AI and Copyright in 2023: In the Courts*, COPYRIGHT ALLIANCE (Jan. 4, 2024), https://copyrightalliance.org/ai-copyright-courts/.

[1543] Procedures and Tentative Rulings, Andersen v. Stability AI, Ltd., No. 3:23-cv-00201-WHO (N.D. Cal. May 7, 2024); *see also* Order, Andersen v. Stability AI, Ltd., No. 3:23-cv-00201-WHO (N.D. Cal. Oct. 30, 2023) (granting motion to dismiss but not dismissing direct copyright infringement claims) and Mem., Thomson Reuters v. Ross Intelligence Inc., No. 1:20-cv-00613-SB (D. Del. Sept. 25, 2023) (denying defendant's motion for summary judgment in a machine-learning case).

[1544] Several of the pending cases include an additional cause of action for violating the Digital Millennium Copyright Act (DMCA) Section 1202 by removing copyright management information (CMI) from the original work. That section prohibits (1) "intentionally remov[ing] or alter[ing] any [CMI] and (2) distributing [CMI] that, or

The fair use doctrine under federal copyright law allows the unlicensed use of works in certain circumstances that would otherwise be protected under copyright law. These circumstances are ones where restrictive application of copyright law would infringe upon free expression or curtail other socially beneficial uses of copyrighted material. Paradigmatic examples of such permitted fair use include criticism or commentary about copyrighted works, news reporting, teaching, scholarship, and research.[1545]

a) The four-factor balancing test to assess fair use

Courts in the US assess fair use by employing a four-factor balancing test, the application of which is fact-specific and, therefore, varies from case to case. The four factors set out in Section 107 of the Copyright Act are:

1. the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes,
2. the nature of the copyrighted work,
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole, and
4. the effect of the use upon the potential market for or value of the copyrighted work.[1546]

Though all statutory factors are relevant and no one factor alone is dispositive,[1547] since the U.S. Supreme Court's landmark 1994 decision *Campbell v. Acuff-Rose*, the cornerstone of the fair use analysis has frequently become the first factor, namely, whether the purpose and character of the defendant's use is "transformative."[1548] If a court decides that the use is transformative, that determination tends to trump any countervailing factors and lead to a finding of fair use. This was affirmed by the Second Circuit U.S. Court of Appeals in 2015 in *Authors Guild v. Google*.[1549] At issue was Google's scanning and digitization of copyrighted books to develop its Google Book Search service, which allowed users to search the full text of any book in Google's database. The Second Circuit ruled that, although Google was a commercial enterprise and was scanning/digitizing books in their entirety, the creation of Google Book Search was highly transformative and included significant restrictions (e.g., how much of a given book's text users

---

a work for which the CMI, has been removed or altered, in each case, with the knowledge that this "will induce, enable, facilitate, or conceal an infringement" of copyright. Plaintiffs argue that training a model on copyrighted materials removes the requisite copyright management information, which constitutes a separate DMCA violation. Fair use is generally not recognized as a defense because DMCA is concerned with the integrity of copyright management information, rather than copyright infringement itself. The aforementioned *Andersen* court dismissed the DMCA claims against all defendants.

[1545] 17 U.S.C. § 107.

[1546] *Id. See also* U.S. Copyright Office, *U.S. Copyright Office Fair Use Index* (last updated Nov. 2023), https://www.copyright.gov/fair-use/.

[1547] Harper & Row Publishers v. Nation Enterprises, 471 U.S. 539, 549 (1985).

[1548] *See, e.g.*, Clark D. Asay et al., *Is Transformative Use Eating the World?*, 61 B.C. L. REV. 907 (2020); Peter Henderson et al., *Foundation Models and Copyright Questions*, STANFORD HAI (Nov. 2023) at 2, https://hai.stanford.edu/sites/default/files/2023-11/Foundation-Models-Copyright.pdf (stating that "the transformativeness factor tends to carry the greatest weight when determining fair use and is heavily emphasized in legal assessments").

[1549] Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

could view). The court ruled that those restrictions prevented Google Book Search from substantially affecting the market for the original books.

The U.S. Supreme Court's 2023 decision in *Andy Warhol Foundation v. Goldsmith,*[1550] while reaffirming earlier fair use cases like *Campbell*, muddied the waters somewhat by placing greater emphasis on the *commercial purpose* of the use (one component of the first factor).[1551] The Supreme Court ruled that the licensing of an image created by Andy Warhol, entitled "Orange Prince" and based on a photograph by Lynn Goldsmith, did not constitute fair use when it was used as a magazine cover. This was because Warhol's unauthorized use of Goldsmith's photo, when used as a magazine cover, served the same commercial purpose as Goldsmith's original photograph. This decision will likely shape how often and to what extent lower courts will consider the impact an allegedly infringing work has on the market for the original work (the fourth factor).[1552]

b) Arguments in favor of and against fair use

No cases alleging infringement in the model training process have yet been fully aired in the courts.[1553] When they are, generative AI companies are likely to contend that the first, third, and fourth factors particularly counsel for a finding of fair use.[1554] On the first, they will argue that their use is transformative – the model has an entirely distinct use and purpose that differs from the original work and does not attempt to mimic the original author's expression.[1555] On the third factor, companies will aver that their AI models usually do not regurgitate works from their training sets; any portion of a protected work that is "made accessible to the public" is limited and not substantial relative to the original work as a whole.[1556] And on the fourth factor, though the use of generative AI may arguably dampen the market for certain works, companies may lean on the policy undergirding the Copyright Act: that copyright exists as a mechanism to incentivize creativity in order to "promote the progress of science and useful arts"[1557] and to benefit the public – not solely as an entitlement or reward for a creator's efforts.[1558]

[1550] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023).

[1551] *Id.* at 531.

[1552] *Id.* at 536 n.12.

[1553] And AI companies have been chary about prematurely unveiling their merits arguments. E.g., Mem. of Law in Support of OpenAI Defendants' Mot. to Dismiss, New York Times Co. v. OpenAI, No. 1:23-cv-11195-SHS (Feb. 26, 2024) (moving to dismiss the direct copyright infringement claims as time barred by the statute of limitations).

[1554] Comment of OpenAI to Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, USPTO Docket No. PTO–C–2019–0038 at 5-8, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf#page=5.

[1555] *Id.* at 5-6.

[1556] *Id.* at 7 (citing *Authors Guild*, supra note 1549, where entire books were copied but only small excerpts were made available). However, researchers have found jailbreaking techniques that allow users to bypass guardrails and get models to output substantial portions of copyrighted works. *See* Peter Henderson et al., *Foundation Models and Fair Use*, arXiv (Mar. 28, 2023), https://arxiv.org/pdf/2303.15715.pdf. The ability of models to output significant portions of copyrighted news articles was also central to the New York Times's complaint against OpenAI. *See* Compl. at 2, New York Times v. Microsoft, No. 1:23-cv-11195 (S.D.N.Y. 2023).

[1557] U.S. Const. art. I, § 8.

[1558] Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2021), at 31 (stating that "[W]e must take into account the public benefits the copying will likely produce.").

Zooming out, some copyright scholars[1559] and AI companies[1560] have argued that the copying of protected works into model training datasets is better characterized as the uncopyrightable learning of ideas, which falls on the unprotected side of the "idea-expression distinction."[1561] Mark Lemley and Bryan Casey argued in a 2021 article that when an AI model is trained, much of what it is doing is learning unprotectable ideas in the form of the general structure and patterns in the giant corpus of works that comprise its training dataset. They suggest that the extent to which copying and training are aimed at extracting unprotectable *ideas* from the works (rather than its protected expression), they should be considered fair use under both the first and second factors.[1562] AI companies, including Google and Anthropic, have made similar arguments in comments submitted to the U.S. Copyright Office.[1563]

Copyright plaintiffs counter that the fair use factors do not support a finding of fair use. For instance, AI companies' scraping and training activities may be transformative but they are still typically commercial in nature, which weighs against fair use.[1564] The process of scraping content from the internet for AI training datasets often involves copying works in their entirety, militating against the third fair use factor. And plaintiffs argue that the models themselves, as tools for generating text and images, could significantly affect the market for plaintiffs' original works.

In light of *Warhol*, it is possible that courts will give greater weight to the commercial character of many generative AI tools and the degree to which those tools—and their outputs—eat into the market for various types of creative works, over their public benefit. This evaluation, like much else in the world of fair use cases, is likely to be highly fact-intensive. For one thing, it will likely involve looking at the model's actual outputs and their degree of similarity to the copyrighted works in question (see section on copyright issues related to AI outputs). As for the AI models themselves, here, too, the analysis could differ based on the works and markets at issue. For example, it is not entirely clear that generative AI models are (or soon will be) able to produce outputs of sufficient sophistication or quality to threaten the market for creative works, like novels or screenplays. By contrast, image generators, like DALL·E and Stable Diffusion, seem to more readily pose a threat to photographers, working artists, and companies (like Getty Images) that have historically supplied commoditized visual works in contexts such as journalism, décor, and graphics.[1565]

[1559] Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743 (2021).

[1560] Google, *Comment Letter on Notice of Inquiry on Copyright and Artificial Intelligence*, at 9 (Oct. 30, 2023), https://www.documentcloud.org/documents/24117935-google; Anthropic, *Comment Letter on Notice of Inquiry on Copyright and Artificial Intelligence*, at 11 (Oct. 30, 2023), https://www.documentcloud.org/documents/24117938-anthropic.

[1561] 17 U.S.C. § 102(b); Google LLC v. Oracle Am., Inc., 593 U.S. 1, 13 (2021) ("[C]opyrights protect expression but not the ideas that lie behind it.") (internal marks omitted).

[1562] Lemley & Casey, *supra* note 1559, at 750.

[1563] Google, *supra* note 1560; Anthropic, *supra* note 1560.

[1564] Commercial use is not decisive, however, as courts have rejected the presumption that such uses are unfair. *See* Campbell v. Acuff–Rose Music, Inc., 510 U.S. 569, 584, 594 (1994) and Authors Guild v. Google, Inc., 804 F.3d 202, 219 (2d Cir. 2015).

[1565] *See, e.g.*, Gian Volpicelli, *The new Luddites: AI comes for the creative class*, POLITICO EU (Feb. 20, 2023), https://www.politico.eu/article/artificial-intelligence-technology-art-regulation-copyright/. Another example of backlash by creatives was the controversy around AI-generated opening credits in Marvel's *Secret Invasion*

The fact-specific nature of fair use cases makes speculation about legal outcomes unwise, even in cases dealing with relatively established fact patterns, such as incorporation of an image or music sample into a new work of the same type. These uncertainties are compounded for generative AI, which is both technologically novel and can be put to myriad different uses. Perhaps all that can be said with much certainty at this stage is that fair use is not guaranteed to attach categorically, and when it applies, courts' fair use analysis is likely to be highly fact-specific and may turn out quite differently from case to case, depending on the specific generative AI tool and plaintiff's facts.[1566]

2) US copyright issues related to AI outputs

In addition to lawsuits targeting the *input* of copyrighted works to train AI models (see above), AI companies may also be the subject of lawsuits alleging copyright infringement based on *specific outputs* of their AI models.

a) Copyright infringement over generative AI outputs

Generative AI model outputs could be seen as derivative works that infringe on copyrighted works, if certain conditions are met. To make a valid legal claim for copyright infringement by an AI output, copyright owners must show that (1) the defendant (presumably the model developer) *actually copied* the copyrighted work, and (2) there is *substantial similarity* between the protected elements of the copyrighted original and the defendant's (allegedly infringing) derivative work.[1567]

i. Actually copied

It is sometimes difficult to prove that one person has copied another person's work directly – and particularly so in the case of generative AI, when the corpus of training materials is not publicly known – and particularly so in the case of generative AI, when the corpus of training materials is not publicly known. Courts typically allow a plaintiff to use circumstantial evidence about the defendant's access to the plaintiff's original work and substantial similarities between the works that are probative of copying. In cases involving generative AI outputs, plaintiffs may seek to establish proof of access by showing that the original work was included in the AI model's training dataset or that it was available on public-facing internet sites that generative AI companies scraped when assembling their training datasets. Similarities probative of copying, meanwhile, could include reproduction of watermarks or other incidental features embedded in the protected work embedded in the protected work, as well as a degree of similarity so high that it is very unlikely to have arisen if the later work were created independently of the earlier (copyrighted) one.

---

television show. Adrian Horton, *Marvel Faces Backlash over AI-Generated Opening Credits*, THE GUARDIAN (June 21, 2023), https://www.theguardian.com/tv-and-radio/2023/jun/21/marvel-ai-generated-credits-backlash.

[1566] Peter Henderson et al., *Foundation Models and Copyright Questions*, STANFORD HAI (Nov. 2023), https://hai.stanford.edu/sites/default/files/2023-11/Foundation-Models-Copyright.pdf at 5-6.

[1567] *Generative Artificial Intelligence and Copyright Law*, CONGRESSIONAL RESEARCH SERVICE (Sept. 29, 2023), https://crsreports.congress.gov/product/pdf/LSB/LSB10922 at 4.

When it comes to the question of access, AI companies' expansive scraping of the internet to assemble giant training datasets seems likely to work against AI companies. Any plaintiff whose works are available on the internet will have a colorable argument that AI companies had access to the work. Furthermore, although many AI companies are quite secretive about the contents of their training datasets, plaintiffs whose litigation gets past an initial motion to dismiss could use discovery to compel disclosure of what works are contained in the defendant's training datasets.

ii) Substantial similarity

Things get more interesting when turning to the second element a plaintiff must show to prove copyright infringement: substantial similarity. The essentially limitless variety of potential generative AI outputs means that there is no general answer for whether these outputs will be deemed substantially similar to original copyrighted works. Rather, substantial similarity will depend on the details of the allegedly infringing output and its degree of similarity to protected elements of the original work. And indeed, this is where plaintiffs in lawsuits against OpenAI and Meta stumbled; they did not satisfactorily allege that ChatGPT and Llama's outputs were substantially similar to their protected works, and so their infringement claims were dismissed (albeit with leave to amend and refile the lawsuit).[1568]

The argument for substantial similarity will be strongest when the AI model is outputting significant, unbroken portions of a copyrighted work. For example, a group of Stanford scholars was able to get ChatGPT to regurgitate large chunks of copyrighted works, including three and a half chapters from *Harry Potter and the Sorcerer's Stone*.[1569] In such situations, the copyright holders could then plausibly claim infringement of their exclusive right to reproduce their work and, perhaps, their exclusive right to distribute. This is not always easy to do without intentional, targeted efforts, which will likely be relevant to the ultimate adjudication of this question. In a response brief to the *New York Times* lawsuit against it, OpenAI characterized such efforts to generate similar or verbatim outputs as "hacking" its products by submitting thousands of "deceptive prompts that blatantly violate OpenAI's terms of use."[1570]

Another situation where copyright holders may sometimes have plausible infringement claims is when a model's output utilizes, as part of a new work, well-defined, distinctive characters or other protected elements from the copyrighted works. For example, if instead of prompting a chatbot to reproduce, verbatim, chapters from a *Harry Potter* novel, a user instead prompts the chatbot to create a new story about

---

[1568] Order, Tremblay v. OpenAI, Inc., No. 23-cv-03416-AMO, 2024 U.S. Dist. LEXIS 24618 (N.D. Cal. Feb. 12, 2024); Order, Kadrey v. Meta Platforms, Inc., No. 3:23-cv-03417-VC (N.D. Cal. Nov. 20, 2023).

[1569] Henderson et al., *Foundation Models and Copyright Questions*, *supra* note 1548, at 8. The authors note that the ability of a model to regurgitate portions of long-form works was likely constrained by the size of a model's context window, which helps explain why they were able to get ChatGPT to regurgitate larger portions of Harry Potter text when using the GPT-4-based version of the chatbot (which has a larger context window). This also suggests that as companies like Anthropic and OpenAI update their models with larger context windows, it may be possible to elicit even larger outputs of copyrighted works unless guardrails are strengthened.

[1570] Mem. of Law in Supp. of OpenAI Defs.' Mot. to Dismiss, N.Y. Times Co v. Microsoft Corp., No. 1:23-cv-11195-SHS (Feb. 26, 2024), at 11.

the same characters—say, a sci-fi crossover involving Harry and his friends traveling to the moon for an adventure. This could infringe on the exclusive right of the copyright holder to create derivative works.[1571]

The trickiest cases are likely to be ones where generative AI outputs are not directly regurgitating substantial portions of existing works or incorporating specific protected content (e.g., characters) but are instead utilizing more general elements, ideas, or styles learned from the training data. Such cases will implicate a range of copyright law principles and limitations. One such limitation is the distinction in copyright law between *ideas* (which can *not* be copyrighted) and particular *expressions of those ideas* (which can be copyrighted). These cases could also test the lines between small building blocks of expression (e.g., individual words or musical notes, short written or musical phrases), genre conventions/tropes (e.g., scènes à faire), or general artistic styles –all of which generally do not enjoy copyright protection.[1572]

The extent to which stylistic elements enjoy copyright protection could become a particularly important issue given the ability of AI models to generate new works that are "in the style of" specific artists or genres but do not closely resemble any particular previous work. User prompts seeking responses "in the style of" famous creators are highly popular, and style alone is not protectable under copyright law (see above). Despite this, some AI companies—perhaps anxious about potential copyright lawsuits—have begun voluntarily implementing guardrails to prevent models from fulfilling prompts asking for outputs in the style of specific artists. For example, OpenAI, in its documentation for DALL·E 3, now states that "DALL·E is designed to decline requests that ask for an image in the style of a living artist. Creators can now also opt their images out from training of our future image generation models."[1573]

One situation where plaintiffs are unlikely to succeed in proving copyright infringement by generative AI outputs is where a model's output is simply summarizing or answering questions about a work without directly quoting the work or taking expressive content from it. For instance, a user may ask a chatbot to list five major life events that a celebrity described in their autobiography. Such outputs, which relay facts or ideas from a written work but not the particular words that the author used to express them, would probably not infringe on the author's copyright since facts and ideas are not protectable under US copyright law.[1574]

One overarching consideration is that plaintiffs' success in holding AI developers liable may depend on how easy or difficult it is to get the model to output the allegedly infringing content. For example, if a model puts out large, verbatim chunks of a copyrighted work in response to vague prompts or only rudimentary prompt engineering (e.g., instructing the model to replace certain letters with similar-looking numbers), the case for holding the AI developer liable may be stronger. (All the more so if the model developer fails to address the problem after being given notice of it by the copyright holder.) By contrast, if the model's

---

[1571] *See* Anderson v. Stallone, No. 87-0592 WDKGX, 1989 WL 206431 (C.D. Cal. Apr. 25, 1989) (holding that a screenwriter's unauthorized script using characters that Sylvester Stallone had created for the *Rocky* movies infringed on Stallone's exclusive right to prepare derivatives works).

[1572] As Lemley & Casey explain, *supra* note 1559 at 778, some recent cases have chipped away at these principles, at least in the context of music. *See* Williams v. Gaye, 895 F.3d 1106 (9th Cir. 2018); *see also* Hall v. Swift, 786 Fed. App'x 711 (9th Cir. 2019).

[1573] *Dall·E 3*, OpenAI (last visited Feb. 23, 2024), https://openai.com/dall-e-3.

[1574] *See* the discussion of the idea-expression distinction in Section 5.3.1.B.

guardrails are robust enough that the model generates copyrighted content only in response to numerous carefully engineered prompts, the case for holding the AI model developer liable is weaker.

This naturally begs the question: When should AI users who write the prompts be held liable? As of the writing of this report, there have been no infringement actions taken against generative AI users for their use of the outputs of models. Many AI developers, including Microsoft,[1575] Google,[1576] OpenAI,[1577] and Anthropic,[1578] have pledged to indemnify certain users, particularly enterprise customers who do not fine-tune or modify the models, against intellectual property claims arising from infringing outputs. While Microsoft and Google have apparently extended this indemnification to all users, OpenAI and Anthropic have limited it to users of their premium or business tiers.[1579]These commitments likely reflect their assessment that the risk of copyright and intellectual property infringement liability is low.

b) Copyrightability of AI-generated content

There are also uncertainties about whether AI generated-content is eligible for copyright protection. Two strands of copyright law and policy may come into tension in answering this question.

The first strand emphasizes that copyright requires a human author. For example, in a much-publicized case from 2018, *Naruto v. Slater*, the Ninth Circuit U.S. Court of Appeals ruled that a selfie taken by a monkey—supposedly without any involvement from the professional photographer who owned the camera—was *not* eligible for copyright protection since the image had no human author.[1580] The requirement of human authorship has been subsequently affirmed for AI-generated works, too, both by a March 2023 guidance document from the U.S. Copyright Office[1581] and an August 2023 federal district court decision, *Thaler v. Perlmutter*.[1582]

[1575] Brad Smith & Hossein Nowbar, *Microsoft announces new Copilot Copyright Commitment for customers*, MICROSOFT (Sept. 7, 2023), https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/.

[1576] Neal Suggs & Phil Venables, *Shared fate: Protecting customers with generative AI indemnification*, GOOGLE CLOUD (Oct. 12, 2023), https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification.

[1577] OpenAI Business terms (last updated Nov. 14, 2023), OPENAI https://openai.com/policies/business-terms/ (Indemnification for business users).

[1578] Anthropic, PBC Commercial Terms of Service (effective Jan. 2024), ANTHROPIC, https://www-cdn.anthropic.com/files/4zrzovbb/website/786ea99408c7b0c14684b6cf4e1b31d34b7a77aa.pdf (indemnification for business users).

[1579] Isabel Gottlieb & Kyle Jahner, *Microsoft Sees Low Risk for Customers in AI Copyright Lawsuits*, BLOOMBERG LAW (Sept. 11, 2023), https://news.bloomberglaw.com/artificial-intelligence/how-risky-is-microsofts-pledge-to-defend-ai-copyright-lawsuits.

[1580] Naruto v. Slater, 888 F.3d 418 (9th Cir. 2018).

[1581] Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16190, 16192 (Mar. 16, 2023) (to be codified at 37 C.F.R. pt. 202).

[1582] Thaler v. Perlmutter, No. 22-1564 (BAH), 2023 WL 5333236 (D.D.C. Aug. 18, 2023).

The U.S. Copyright Office's guidance highlighted its longstanding application of the human authorship requirement. It stated that, when the "traditional elements of authorship were produced by a machine, the work lacks human authorship and the Office will not register it."[1583] The guidance went on to state that while some works containing AI-generated material are protectable, "when an AI technology receives solely a prompt from a human and produces complex written, visual, or musical works in response, the 'traditional elements of authorship' are determined and executed by the technology—not the human user." In other words, "[w]hen an AI technology determines the expressive elements of its output, the generated material is not the product of human authorship."[1584] That material is not protected by copyright. However, technological tools like AI can be part of a human author's creative process.[1585]

The August 2023 federal court decision in *Thaler v. Perlmutter* involved a plaintiff, Steven Thaler, who claimed that his AI system had created a digital image without any human prompting or other guidance. Thaler sought a copyright of the image in the name of his AI model and then sought to transfer the copyright to himself as owner of the model. The court found that, since there was no human authorship, the work was ineligible for copyright.[1586] But given the unusual nature of Thaler's claim of an entirely autonomous AI creation, the court's language about the need for a "guiding human hand" could leave open the possibility that other products of generative AI—namely, those where human guidance plays more of a role—are copyrightable.[1587]

A second strand of copyright law and policy has long acknowledged that novel technologies *can* open up new or different creative processes that still constitute human authorship. The seminal U.S. Supreme Court decision here is the 1884 *Burrow-Giles Lithographic Co. v. Sarony*, which dealt with the question of whether images created using the then-new technology of photography could be copyrighted.[1588] The Supreme Court said yes. Therefore, it is possible that, as generative AI becomes an accepted, widespread part of the creative process—much like photography, digital image manipulation, and other once-new technologies—it will come to be seen as just another tool humans use to create expressive works, the involvement of which does not preclude copyrightability. Greater understanding of the effort, expertise, and creativity that often goes into prompting AI models to get a desired output may lead the Copyright Office to relax its stance that prompting alone is not a sufficient human contribution for authorship of the

[1583] Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, *supra* note 1581.

[1584] *Id.* ("Based on the Office's understanding of the generative AI technologies currently available, users do not exercise ultimate creative control over how such systems interpret prompts and generate material. Instead, these prompts function more like instructions to a commissioned artist—they identify what the prompter wishes to have depicted, but the machine determines how those instructions are implemented in its output.")

[1585] *Id.*

[1586] *Thaler*, 2023 WL 5333236, at *1. Thaler's similar efforts under a different IP regime, patent law, have met a similar fate. *See* Thaler v. Vidal, 43 F.4th 1207 (Fed. Cir. 2022), *cert. denied*, 143 S. Ct. 1783 (2023).

[1587] Paul Goldstein, a leading copyright scholar, makes this point in a recent interview. *See* Paul Goldstein, *The Writers' Strike Four Months In: Stanford's Paul Goldstein on Artificial Intelligence and the Creative Process*, STANFORD LAW SCHOOL (Sept. 5, 2023), https://law.stanford.edu/2023/09/05/the-writers-strike-four-months-in-stanfords-paul-goldstein-on-artificial-intelligence-and-the-creative-process/.

[1588] Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884).

resulting output.[1589] This issue is likely to be evaluated further by the Copyright Office, as it presently considers comments in response to its August 2023 "Notice of Inquiry on Copyright and Artificial Intelligence."[1590]

3) Patentability questions

A similar set of questions has arisen in the context of another intellectual property regime: patent law. Stephen Thaler is again an important test case plaintiff in early rulings on these issues. This time, Thaler is claiming that a different AI system he created generated two patentable inventions. In his applications to the U.S. Patent and Trademark Office (USPTO), Thaler listed his AI system as the sole inventor of two inventions. The USPTO denied the patent applications on the grounds that they failed to list a valid (human) inventor. The Office stated that, "[t]o the extent the petitioner argues that an 'inventor' could be construed to cover machines, the patent statutes preclude such a broad interpretation."[1591] Thaler challenged the Office's decision in court. However, both the district and appellate courts reached the same conclusion as the USPTO: that an "inventor," as defined in the Patent Act, is limited to natural persons (i.e., human beings).[1592]

Much like Thaler's copyright case, his patent case, *Thaler v. Vidal*, sought (and failed) to establish that an AI system itself can be recognized as the creator of intellectual property under US law. More interesting and practically relevant are questions about the patentability implications of invention processes involving both human and AI contributions, something that is already happening in areas such as AI-assisted drug discovery. The appellate court in *Thaler v. Vidal* limited its ruling to the issue before it: Thaler's contention that an AI system can be an "inventor" under the Patent Act. It expressly declined to address "the question of whether inventions made by human beings with the assistance of AI are eligible for patent protection."[1593] But for the same reason that the court rejected Thaler's effort to have an AI system recognized as the sole inventor of a patentable invention—the fact that the "inventor" was not a human individual—it also cannot be a joint inventor alongside humans.[1594] Thus, a key issue in future USPTO and court decisions on the patentability of AI-assisted inventions may be whether the humans involved have made the sorts of

---

[1589] Indeed, if generative AI shifts the locus of creativity toward devising prompts and away from crafting the expressive work itself, this could put significant strain on long-standing legal doctrines and incentive structures that underpin US copyright law. *See* Mark A. Lemley, *How Generative AI Will Turn Copyright on its Head*, COLUM. SCI. & TECH. L. REV. (forthcoming 2024).

[1590] U.S. Copyright Office, *Copyright Office Issues Notice of Inquiry on Copyright and Artificial Intelligence* (Aug. 30, 2023), https://www.copyright.gov/newsnet/2023/1017.html;
Artificial Intelligence and Copyright, 88 Fed. Reg. 59, 942 (Aug. 30, 2023), https://www.govinfo.gov/content/pkg/FR-2023-08-30/pdf/2023-18624.pdf; *see also* Ltr. from US Copyright Office to Senators (Feb. 23, 2024), https://copyright.gov/laws/hearings/USCO-Letter-on-AI-and-Copyright-Initiative-Update.pdf.

[1591] *Ex parte* Flashpoint IP, No. 50567-3-01-US (July 29, 2019) at 4, https://www.uspto.gov/sites/default/files/documents/16524350_22apr2020.pdf.

[1592] Thaler v. Vidal, 43 F.4th 1207, 1208 (Fed. Cir. 2022), *cert. denied*, 143 S. Ct. 1783 (2023).

[1593] *Id.* at 1213.

[1594] *Id.* at 1211.

contributions—specifically, conceiving of the invention or reducing it to practice[1595]—required for *them* to count as inventors under US patent law.[1596]

In early 2024, the Patent Office published guidance pursuant to White House Executive Order 14110 and in the wake of *Thaler v. Vidal*, affirming that AI-assisted inventions are patentable if human contributions are significant.[1597] The Patent Office's guidance and other patentability-relevant considerations will doubtless shape how sophisticated actors, like pharmaceutical companies, design their AI-assisted development pipelines, as well as how they frame the involvement of AI tools when applying to patent any resulting inventions.

5.3.1.C Liability for machine-generated content

To what extent might companies that develop generative AI models and systems be held liable under US law for harms caused by their tools? Companies like OpenAI[1598] and Google[1599] include disclaimers about their chatbot systems to alert users about the potential for inaccurate or misleading outputs. But it is unclear to what extent such contractual disclaimers or admonitions may thwart liability.

To date, litigation concerning outputs generated by models has been limited, with a few notable exceptions involving defamation cases. If generative AI systems produce false statements, such as unfounded accusations of misconduct or criminal convictions, the victims of such reputational harms may seek legal recourse under defamation laws. One prominent defamation case, Walters v. OpenAI, has survived a motion to dismiss, meaning that at least one court recognizes a plaintiff's legal theory for "defamation by chatbot" has prima facie validity.[1600] In that case, plaintiff Mark Walters has alleged that OpenAI's ChatGPT hallucinated false outputs about him that defamed him –even though the user who generated those outputs did so privately and did not disseminate them to anyone except Walters.[1601] OpenAI argued that there was no publication of the defamatory output, that the plaintiff is a public figure, and that there was no actual

[1595] *See* Dana-Farber Cancer Institute v. Ono Pharm. Co., 964 F.3d 1365, 1371 (Fed. Cir. 2020).

[1596] For an accessible discussion of this and related issues, *see* Ben Hsing, *Artificial Intelligence in Drug Development: Patent Considerations*, IPWATCHDOG (Sept. 25, 2023), https://ipwatchdog.com/2023/09/25/artificial-intelligence-drug-development-patent-considerations/id=167125/.

[1597] Dep't of Commerce, Patent and Trademark Office, *Inventorship Guidance for AI-Assisted Inventions*, 89 Fed. Reg. 10043 (proposed Feb. 13, 2024).

[1598] *Terms of Use*, OPENAI (last visited Mar. 16, 2024), https://openai.com/policies/terms-of-use.

[1599] *Google Privacy and Terms*, GOOGLE (last visited Apr. 6, 2024), https://policies.google.com/terms; *Generative AI Additional Terms of Service*, GOOGLE (last visited Apr. 6, 2024), https://policies.google.com/terms/generative-ai ("Use discretion before relying on, publishing, or otherwise using content provided by the Services.").

[1600] Order, Walters v. OpenAI, LLC, No. 23-A-04860-2 (Ga. Super. Ct. Jan. 11, 2024).

[1601] Compl. at ¶¶ 33–37, Walters v. OpenAI, LLC, No. 23-A-04860-2 (Ga. Super. Ct. June 5, 2023); Eugene Volokh, Court Lets First AI Libel Case Go Forward (Jan. 17, 2024), https://reason.com/volokh/2024/01/17/court-lets-first-ai-libel-case-go-forward/.

malice (i.e., knowledge that the statements were false or reckless disregard for their truth or falsity).[1602] Nevertheless, a Georgia state court judge allowed the case to proceed in January 2024.

A second pending case involves a plaintiff who has alleged that searching his own name on Microsoft's Bing search engine returned an AI-generated summary that commingled facts about him with facts about a different individual with a similar name who once pleaded guilty to a serious crime.[1603] More defamation cases like this for machine-generated content are sure to follow.

Beyond defamation, false information from chatbots may lead to other types of real-world harm. For example, inaccurate medical or legal information could lead AI users, relying in good faith on the model's guidance, to make choices or do things that cause physical or legal harm to themselves or others. Some scholars contend that companies could be liable regardless, if, for example, they offered the recipe for a poisonous concoction, even when so prompted by a user[1604] – which may be one reason many chatbots decline to respond to such requests and why many developers strive to add safeguards and achieve ethical alignment with their models.[1605] However, the question of civil liability remains open and may require a rethinking of legal doctrine.[1606] The following developments will address only a few outstanding issues.

1) Could AI-generated content be considered constitutionally protected free speech?

The First Amendment of the U.S. Constitution states that "Congress shall make no law... abridging the freedom of speech." This is often referred to as the freedom of "expression" and encompasses myriad forms of "speech": political protest, burning a flag, publishing news stories, access to books, displaying artwork, erecting a cross, etc. Leading legal scholars[1607] have argued that the First Amendment's guarantee of free expression could be a significant barrier to attempts by US government entities to restrict the outputs of generative AI models. Some argue that the First Amendment should also protect the rights of users and others to receive AI model outputs.[1608]

This specific and not-yet-resolved legal question will likely receive much attention if the current debates over the government's role in moderating content on social media spreads to generative AI. However, even if AI model outputs do enjoy First Amendment protection, there are well-known exceptions to that

---

[1602] Def.'s Mem. Supp. Mot. to Dismiss, Walters v. OpenAI, LLC, No. 23-A-04860-2 (Ga. Super. Ct. Nov. 1, 2023).

[1603] Compl., Battle v. Microsoft Corp., No. 1:23-cv-01822-JRR (D. Md. Jul. 7, 2023).

[1604] Ephrat Livni et al., *Who is Liable for A.I. Creations*, N.Y. TIMES (June 3, 2023), https://www.nytimes.com/2023/06/03/business/who-is-liable-for-ai-creations.html.

[1605] *What is AI Alignment?*, IBM RSCH. BLOG (Nov. 8, 2023), https://research.ibm.com/blog/what-is-alignment-ai.

[1606] *See* Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 5 (2019), https://lawreview.uchicago.edu/print-archive/remedies-robots.

[1607] Eugene Volokh et al., *Freedom of Speech and AI Output*, 3 J. FREE SPEECH L. 651 (2023); Cass R. Sunstein, *Artificial Intelligence and the First Amendment*, (Apr. 28, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4431251.

[1608] Eugene Volokh et al., *Freedom of Speech and AI Output*, 3 J. FREE SPEECH L. 651, 655 (2023).

protection, and freedom of speech does not mean that the AI companies or users are immune from any liability. As discussed, they can still potentially be held legally responsible for defamatory speech or speech that forms part of a criminal act, e.g., soliciting another person to commit a crime.[1609]

2) Is AI-generated content covered by Section 230 CDA?

Section 230 of the Communications Decency Act of 1996 shields providers or users of "interactive computer services" from being held liable for certain kinds of unlawful content posted on those services by a third-party.[1610] The core of the liability shield is Section 230(c)(1), which states: "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by *another* information content provider."[1611] Courts have broken this statutory language into a three-element test to determine whether content is protected: The defendant must be (1) a provider or user of an interactive computer service, and the plaintiff's lawsuit must be (2) seeking to hold the defendant liable as the publisher or speaker of (3) content from another information content provider.[1612] Paradigmatically, Section 230 protects services like Facebook, Google, and Amazon from being held liable for content uploaded or produced by users and other information content providers.[1613]

In generative AI cases, the analysis for the first two elements will often be relatively straightforward. The statute's definition of an "interactive computer service" is extremely broad, encompassing almost any digital service that utilizes online functionality.[1614] Online chatbots and other tools built atop generative AI models are all but certain to count as interactive computer services.[1615] Many legal claims that may arise in the context of generative AI—such as civil suits for reputational harm or personal injury as a result of false model outputs—will thus turn on the third element. They will turn on whether the claim would hold a chatbot (or, more accurately, the company that developed or deployed the chatbot) liable for content from *another* information content provider.[1616]

At first blush, it may seem self-evident that generative AI models are (as the term suggests) *generating* their own content, rather than just providing a forum for a third party's material. This is certainly the view of the

---

[1609] For discussion of the nuances of these and other legal regimes' interaction with the First Amendment, *see id.*

[1610] 47 U.S.C. § 230.

[1611] 47 U.S.C. § 230(c)(1) (emphasis added).

[1612] Matt Perault, *Section 230 Won't Protect ChatGPT*, 3 J. FREE SPEECH L. 363, 364–65 (2023).

[1613] Peter J. Benson & Valerie C. Brannon, CONG. RESEARCH SERV., LSB11097, *Section 230 Immunity and Generative Artificial Intelligence* 2 (Dec. 28, 2023), https://crsreports.congress.gov/product/pdf/LSB/LSB11097.

[1614] Specifically, 47 U.S.C. § 230(f)(2) defines an "interactive computer service" as "any information service, system, or access software provider that provides or enables computer access by multiple users to a computer server, including specifically a service or system that provides access to the Internet and such systems operated or services offered by libraries or educational institutions."

[1615] Peter Henderson et al., *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589, 621 n.110 (2023). The authors also note that a model running locally on a user's device would be less obviously covered, but that it would be technically trivial for companies to add some token online functionality if they want to ensure that the model meets Section 230's definition of an "interactive computer service."

[1616] Courts have, however, found that some product liability claims against online platforms are not attempting to hold the defendant as a publisher/speaker, and therefore are not barred by Section 230. *See, e.g.*, Maynard v. Snapchat, 870 S.E.2d 739 (Ga. Sup. Ct. 2022).

drafters of Section 230, U.S. Senator Ron Wyden and former U.S. Representative Chris Cox. Both said they think that the law's liability shield should not apply to generative AI.[1617] During oral arguments in *Gonzalez v. Google*, one Supreme Court member, Justice Neil Gorsuch, telegraphed his assent that Section 230 protection may not extend to generative AI.[1618] His position appears to be shared by some legal scholars who have considered the matter.[1619]

However, the counterargument is that AI models do not generate content automatically; they require user input and prompting. The user may be seen as the speaker/publisher and as the sole "information content provider" under Section 230. Moreover, there is considerable nuance in how the third element of the test gets resolved, as there are a variety of ways that generative AI models are designed to respond to queries, with different relationships between third-party content and a model's outputs. In other words, because generative AI "operate[s] on something like a spectrum between a retrieval search engine (more likely to be covered by Section 230) and a creative engine (less likely to be covered)," courts may not land on a single per se rule for whether Section 230 protection applies categorically or not.[1620] The details matter and may lead to different outcomes in different cases, or even different applications of the same model or product, depending on the facts.[1621]

No court has yet ruled on the validity of a Section 230 defense in the generative AI context. In the search engine context, though, courts have found that, even when the search engine's generation of a summary or snippets of search results creates content that is technically new (i.e., content that does not appear verbatim in the source webpage), the summary is still considered fully derived from third-party content and, therefore, covered by Section 230.[1622] Some industry advocates[1623] and academics[1624] have argued that such holdings should extend to generative AI outputs. They contend that, even when outputs seem novel or creative, they are still ultimately dependent on third-party content from training data and user prompts. This view finds further support with the limited case law involving auto-completed or suggested search terms.

[1617] Cristiano Lima-Strong, *AI Chatbots Won't Enjoy Tech's Legal Shield, Section 230 Authors Say*, WASH. POST (Mar. 17, 2023), https://www.washingtonpost.com/politics/2023/03/17/ai-chatbots-wont-enjoy-techs-legal-shield-section-230-authors-say/. Senators Richard Blumenthal and Josh Hawley have proposed a bill that would amend the text of Section 230 to explicitly state that it does not apply to generative AI. Press Release, Josh Hawley, *Senator, Hawley, Blumenthal Introduce Bipartisan Legislation to Protect Consumers and Deny AI Companies Section 230 Immunity* (June 14, 2023), https://www.hawley.senate.gov/hawley-blumenthal-introduce-bipartisan-legislation-protect-consumers-and-deny-ai-companies-section.

[1618] Tr. of Oral Arg. at 51, Gonzalez v. Google LLC, 598 U.S. 617 (2022) (No. 21–1333).

[1619] *See, e.g.*, Peter Henderson et al., *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589, 622 (2023); Eugene Volokh, *Large Libel Models*, 3 J. FREE SPEECH L. 489, 494 (2023); Matt Perault, *Section 230 Won't Protect ChatGPT*, 3 J. FREE SPEECH L. 363, 365 (2023).

[1620] Benson & Brannon, *supra* note 1613 at 3; *see also* Henderson et al., *Where's the Liability in Harmful AI Speech?*, *supra* note 1615, at 622.

[1621] Henderson et al., *Where's the Liability in Harmful AI Speech?*, *supra* note 1620, at 622.

[1622] *See, e.g.*, O'Kroley v. Fastcase, Inc., 831 F.3d 352, 355 (6th Cir. 2016); Maughan v. Google Tech., Inc., 49 Cal. Rptr. 3d 861 (Cal. Ct. App. 2006).

[1623] For example, Jess Miers of the tech industry group Chamber of Progress. *See* Jess Miers, *Yes, Section 230 Should Protect ChatGPT and Other Generative AI Tools*, TECHDIRT (Mar. 17, 2023), https://www.techdirt.com/2023/03/17/yes-section-230-should-protect-chatgpt-and-others-generative-ai-tools/.

[1624] Derek Bambauer & Mihai Surdeanu, *Authorbots*, 3 J. FREE SPEECH L. 375 (2023), https://www.journaloffreespeechlaw.org/bambauersurdeanu.pdf.

Two district courts rejected defamation claims against search engines for allegedly defamatory auto-generated or suggested search terms, because they merely indicated other websites have connected the ideas, not the search engine itself.[1625] Similarly, an appellate court has held that Facebook's predictive algorithms that merely organize and arrange third-party content does not make Facebook a publisher of content.[1626] It could be argued that generative AI operates in a similar way, even if it does appear to produce new content of its own.

This argument becomes strained on the other end of the spectrum, with model hallucinations seeming less likely to be protected, as they represent creative, brand-new (and false) text that no other party has ever written. When a model makes up falsehoods out of whole cloth or draws incorrect inferences from data[1627] in ways that are highly damaging to people's reputations, the AI model more likely has "materially contribut[ed]" to what makes the content legally actionable, and the Section 230 liability shield will not likely apply.[1628] In sum, Section 230 immunity is unlikely to generally protect AI generated content and may not shield against lawsuits, especially where the model hallucinates false and damaging information about a real person.[1629]

3) How may the rules of civil liability apply to AI developers?

Even if AI companies do not enjoy Section 230 protection for the outputs of their generative models, there are further requirements that must be met to establish substantive liability under another legal regime. Take defamation: That tort's standard elements are (1) the publication of (2) a false statement that (3) causes harm to a person's reputation with (4) a culpable mental state.[1630] Elements (2) and (3), falsity and reputational harm, are relatively easy to establish in many cases, given that chatbot outputs have falsely linked real people to personal and legal misconduct.[1631] And in defamation law, "publication" refers to any

[1625] Benson & Brannon, *supra* note 1613, at 4 (citing Stayart v. Google Inc., 783 F. Supp. 2d 1055 (E.D. Wis. 2011) and Obado v. Magedson, No. 13-2382 JAP, 2014 WL 3778261 (D.N.J. July 31, 2014)).

[1626] *Id.;* Force v. Facebook, Inc., 934 F. 3d 53, 66 (2nd Cir. 2019).

[1627] For example, conflating a businessman with a similarly (but not identically) named terrorist and generating a biographical paragraph that attributes the latter's crimes to the former. *See* Eugene Volokh, *Large Libel Models*, *supra* note 1619; Eugene Volokh, *New Lawsuit Against Bing Based on Allegedly AI-Hallucinated Libelous Statements*, THE VOLOKH CONSPIRACY (Jul. 13, 2023), https://reason.com/volokh/2023/07/13/new-lawsuit-against-bing-based-on-allegedly-ai-hallucinated-libelous-statements/.

[1628] Fair Hous. Coun. of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1168 (9th Cir. 2008). For discussion, *see* Volokh, *Large Libel Models*, *supra* note 1619 *at* 495–98.

[1629] Benson & Brannon, *supra* 1613 at 4.

[1630] The exact mental state requirement varies, typically between either "actual malice" or negligence depending on whether the plaintiff is a public figure or private citizen.

[1631] *See, e.g.*, Volokh, *Large Libel Models*, *supra* 1619 at 555–57; Pranshu Verma & Will Oremus, *ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused*, WASH. POST (Apr. 5, 2023), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/; Tiffany Hsu, *What Can You Do When A.I. Lies About You?*, N.Y. TIMES (Aug. 3, 2023), https://www.nytimes.com/2023/08/03/business/media/ai-defamation-lies-accuracy.html.

communication of the content to a third party (i.e., someone other than the person being defamed), rather than the everyday meaning, i.e., dissemination to a broad audience.[1632]

Therefore, establishing a culpable mental state, such as negligence, is likely to be the main obstacle to defamation claims (and other legal claims with similar mental state requirements). The best facial argument against liability is that, since generative AI models do not have minds, they cannot have mental states, including mental states that are required for torts like defamation.[1633] Of course, as with lawsuits involving other products and services, plaintiffs would be suing the developer of the chatbot (a company) and would likely seek to establish that the company or its employees were negligent or otherwise had the culpable mental state –not the chatbot itself. Establishing such culpability with regard to any particular output of the model is likely to be a difficult and fact-intensive task,[1634] though plaintiffs' prospects may improve if they can show that the company was on notice about its model's false outputs.[1635]

More pragmatically, even should liability be established, another burden awaits plaintiffs – proving damages resulted from the false outputs. If outputs are not disseminated broadly or viewed by many others, then the defamed individuals may have trouble proving they suffered significant damages. This compounds the general difficulty in quantifying damages for reputational and other nonpecuniary harms.[1636]

4) Do product liability rules apply?

Though defamation is the most cited example of civil liability claim that could arise out of using generative AI, product liability could theoretically capture any harm caused by the technology itself. That might include things like harmful instructions or illicit or erroneous advice (e.g., a medical diagnostic system that fails to detect a disease).[1637] Some scholars have suggested that product liability law could be both legally and conceptually useful as a way to deal with chatbot-caused harms like defamation.[1638] Although there may be obstacles to directly applying product liability law to reputational harms,[1639] such an approach may offer certain advantages as a framework for thinking about the law and policy of generative AI liability.

[1632] *See* Volokh, *Large Libel Models*, *supra* note 1619 *at* 504–05 (2023).

[1633] *See* Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. FREE SPEECH L. 389, 399–401 (2023); Peter Henderson et al., *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589, 640–41 (2023). *See generally* Mark Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1311 (2019).

[1634] Henderson et al., *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589, 641 (2023).

[1635] Volokh, *supra* note 1619 at 516–17.

[1636] Some jurisdictions operate on a "presumed damages" rule where plaintiffs can still collect nominal damages without proof of impairment of reputation. But others require a showing of actual injury. *See* 4 Modern Tort Law: Liability and Litigation § 35:32 (May 2023 update).

[1637] Brown, *supra* note 1633, at 396; John Villasenor, *Products Liability Law as a Way to Address AI Harms*, BROOKINGS INSTITUTE (Oct. 31, 2019), https://www.brookings.edu/articles/products-liability-law-as-a-way-to-address-ai-harms/.

[1638] Brown, *supra* note 1633. *See also* Eugene Volokh, *Large Libel Models*, supra note ; Christopher Mims, *The AI Industry Is Steaming Toward a Legal Iceberg*, WALL ST. J. (Mar. 29, 2024), https://www.wsj.com/tech/ai/the-ai-industry-is-steaming-toward-a-legal-iceberg-5d9a6ac1.

[1639] Volokh, *supra* note 1619, at 524–25.

Product liability is a form of tort law that places legal responsibility on product manufacturers and distributors if products they produce or sell are defective. Product liability derives from the common law and varies by state, but one type of defect that is generally cognizable across jurisdictions is a design defect. Defective design causes of action go directly to analysis of whether the company was negligent or otherwise legally culpable in the design of the product, rendering the product unsafe.[1640] It may be possible to show that the design of a chatbot was defective in various ways –such as using a flawed dataset or having inadequate guardrails to prevent false outputs. Plaintiffs could argue that those flaws caused the AI model's production of harmful outputs and were reasonably foreseeable.[1641]

There is no equivalent in the US to the EU's Product Liability Directive (*see above section 5.1.3.A*) , and no US court has yet determined that AI models are products. In fact, in only a few cases have courts held that software can be a product for purposes of product liability,[1642] and that proposition, even if it has backing from prominent legal treatises,[1643] may be debated. Software as a product is likely to be strained further in AI, because, at different points in the AI supply chain (and the more bespoke the offering), what is being provided looks less like a product and more like a service, which would put it outside the reach of product liability claims. In addition, harms, such as damage to one's reputation or business, may not be cognizable under traditional product liability law when there is no accompanying physical damage to persons or property.[1644] Some scholars have argued that this is not an insurmountable obstacle even under current legal doctrine, but that remains to be seen.[1645]

These or other efforts to extend product liability laws to generative AI tools could, however, suffer from First Amendment infirmities. Applying product liability to speech products, like AI chatbots, could restrict expression in a way that impinges upon the countervailing protections of the First Amendment. Exactly how such First Amendment challenges play out is likely to depend on the details of the laws and AI tools at issue in each case. But over the past few decades, courts have significantly expanded the range of situations where they deem the First Amendment to limit the scope of tort liability.[1646] It seems likely that judges will, at a minimum, take seriously the possibility that the First Amendment limits the extent to which some product liability regimes can be applied to generative AI.

Overall, it appears unlikely that any of the few questions raised here may receive a definitive answer at this time. And it is probable that additional questions will emerge in the near future. For instance, one might

---

[1640] Cause of Action for Personal Injury Caused by Defective Design of Product at Section 3, 13 Causes of Action 595 (last updated Feb. 2024). *See* Brown, *supra* note 1633, at 410–14 (2023); Eugene Volokh, *Large Libel Models*, 3 J. FREE SPEECH L. 489, 523–26 (2023).

[1641] Brown, *supra* note 1633, at 411–12 (2023); Eugene Volokh, *Large Libel Models*, supra note 1619. They would also need to argue there existed a feasible alternative design for the model.

[1642] Brown, *supra* note 1633, at 404; *see also* Brenda Leong and Jey Kumarasamy, *Third-party liability and product liability for AI systems*, IAPP (July, 26, 2023), https://iapp.org/news/a/third-party-liability-and-product-liability-for-ai-systems/.

[1643] *Id.* at 405 (citing Restatement (Third) of Torts: Prod. Liab. § 19 (1998)).

[1644] Brown, *supra* note 1633, at 407.

[1645] *Id.* at 406–09.

[1646] *See* Kenneth S. Abraham & G. Edward White, *First Amendment Imperialism and the Constitutionalization of Tort Liability*, 98 TEX. L. REV. 814 (2020).

question whether plaintiffs could successfully argue that certain voluntary commitments made by the leading AI companies and discussed below (*see section 5.3.2.B)* establish a standard of care or have other legal consequences.

5.3.2. US federal regulatory initiatives

The US, by contrast to the EU, lacks a digital regulatory infrastructure that can be adapted to generative AI. Despite a great deal of talk in Washington in recent years about the need for a comprehensive federal consumer privacy law and updated competition laws fit for the digital age, no such legislation has been passed. There are reasons to think that the US may opt for a less aggressive approach to AI regulation than the EU. Historically, all federal tech regulation has stalled in recent decades.[1647] This has been due to a smorgasbord of concerns about impeding innovation through overregulation, a lack of technological expertise among elected officials and regulators, and partisan gridlock in Congress.

So far, the most comprehensive and high-profile US government action on generative AI has been President Joe Biden's "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," issued on October 30, 2023.[1648] The executive order is an important step in its mobilization of government resources and in the message it sends about the attention given to AI at the highest echelons of American government. However, the executive order does not establish a legally binding regulatory regime for the private sector to follow. Binding federal regulations must await new legislation from Congress and/or formal action by regulatory agencies under their existing statutory authority.

Despite the limited progress toward comprehensive, legally binding regulation, the US government's preliminary activity around generative AI does illuminate some broad policy priorities (and tensions) that could eventually shape such regulation. Federal government initiatives on generative AI can be divided into three broad areas: actions by existing federal agencies under existing authority, the Biden administration's strategy, and proposals for future legislation.

5.3.2.A Action by existing federal agencies under existing authority

Even without new laws or agencies specifically focused on regulating generative AI, these models (and the companies that provide or use them) could still be subject to regulation under existing legal regimes. The federal government has a variety of agencies with statutory mandates to regulate particular practices or oversee sectors – though the latitude granted within that mandate has shrunk considerably in the wake of the Supreme Court's watershed decision in Loper Bright Enterprises et al. v. Raimondo (2024), at least with respect to formal rulemaking actions. As AI has rapidly grown in political and economic importance, many of these agencies have begun taking informal action, including by opening investigations, in effect asserting that their existing powers apply to AI companies/tools in the sectors that they already regulate.

[1647] Ian Prasad Philbrick, *The U.S. Regulates Cars, Radio and TV. When Will It Regulate A.I.?*, N.Y. TIMES (Aug. 24, 2023), https://www.nytimes.com/2023/08/24/upshot/artificial-intelligence-regulation.html.

[1648] Exec. Order No. 14110, Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

1) Federal Trade Commission

One agency that is likely to play a leading role in the regulation of generative AI is the Federal Trade Commission (FTC). The FTC is an independent agency led by five presidentially appointed commissioners, with a dual mandate of consumer protection and competition law (antitrust). Current FTC Chair Lina Khan has attempted to exert influence over generative AI early in its life cycle. In a 2023 guest essay for the *New York Times*, Khan argued that the FTC must ensure that "history doesn't repeat itself" with AI as it did with the advent of social media and Web 2.0 in the mid-2000s, when regulation and enforcement lagged the development of those emerging technologies.[1649] Even in the absence of new, specific legal authority to regulate AI, she and other leaders at the FTC contend that the agency's existing powers can be used to protect consumers and competition from harms caused by generative AI.[1650] Two primary sources of these powers are the FTC Act (for consumer protection)[1651] and the Clayton Act (for competition/antitrust).[1652] Though both statutes are over 100 years old and are not directly applicable to AI, the FTC has not hesitated to marshal these powers and use them to investigate AI-related products and services.[1653]

a) The competition concern

FTC staff has highlighted competition as among its chief concerns surrounding generative AI . The FTC observes that the control that large, established tech companies have over three key resources needed to develop cutting-edge generative AI models (data, compute, and talented workers) could stifle competition in the industry (*see section 3.4.1*).[1654] Competition concerns have the potential to intersect with other legal issues beyond the FTC's traditional remit: In comments submitted to the U.S. Copyright Office's call for

---

[1649] Lina M. Khan, *Lina Khan: We Must Regulate A.I. Here's How*., N.Y. TIMES (May 3, 2023), https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html.

[1650] *See* Alvaro M. Bedoya, Comm'r, FTC, Prepared Remarks before the International Association of Privacy Professionals, Early Thoughts on Generative AI, at 15–16 (Apr. 5, 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/Early-Thoughts-on-Generative-AI-FINAL-WITH-IMAGES.pdf; Samuel Levine, Dir. of Bureau of Consumer Prot., FTC, Believing in the FTC, Remarks at Harvard Law School (Apr. 1, 2023), at 8–10, https://www.ftc.gov/system/files/ftc_gov/pdf/Remarks-to-JOLT-4-1-2023.pdf.

[1651] *What the FTC Does*, FTC (last visited Apr. 14, 2024), https://www.ftc.gov/news-events/media-resources/what-ftc-does; Federal Trade Commission Act, 15 U.S.C. §§ 41–58.

[1652] *The Antitrust Laws*, FTC (last visited Apr. 24, 2024), https://www.ftc.gov/advice-guidance/competition-guidance/guide-antitrust-laws/antitrust-laws .

[1653] Press Release, FTC, FTC Authorizes Compulsory Process for AI-related Products and Services (Nov. 21, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/11/ftc-authorizes-compulsory-process-ai-related-products-services; *see also A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority*, FTC (May 2021), https://www.ftc.gov/about-ftc/mission/enforcement-authority.

[1654] Staff in the Bureau of Competition & Office of Tech., FTC, Generative AI Raises Competition Concerns (June 29, 2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns; Lina M. Khan, Chair, FTC, Remarks to Stanford Institute for Economic Policy Research (Nov. 2, 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/khan-remarks-stanford.pdf.

input on AI and copyright, the FTC highlighted concerns about large incumbent tech companies' control over data and computing resources.[1655]

In January 2024, the FTC launched an inquiry into the competitive implications of the three investments cum partnerships between large tech companies Alphabet, Microsoft and Amazon with leading AI startups OpenAI and Anthropic. There is no allegation of any legal violation; the purpose of the inquiry is to produce a study on the competitive impact of these tie-up arrangements, which bundle traditional investment with cloud service provision.[1656] Finally, in early June 2024, the FTC commenced a probe of Microsoft's structuring of a March deal with Inflection AI, whereby Microsoft hired nearly all of Inflection AI's employees and licensed its intellectual property – in effect consummating an acquisition. The probe investigates whether Microsoft structured the transaction to evade antitrust scrutiny, which an outright acquisition would have otherwise faced.[1657]

b) The transparency concern

FTC Commissioner Alvaro Bedoya has highlighted, as a key source of potential risk and harm, the unpredictability and lack of transparency of generative AI models—even to those who create them.[1658] He warned that a product being unpredictable is generally not a defense against legal/enforcement actions resulting from harm that a product causes.[1659] Commissioner Bedoya has criticized OpenAI's technical report accompanying the release of GPT-4, citing its lack of transparency on numerous key features of the model and how it was developed.[1660] In July 2023, the FTC opened a different inquiry[1661] into OpenAI, the overarching subjects of which were whether OpenAI has (1) engaged in unfair or deceptive data or security practices and/or (2) engaged in unfair or deceptive practices relating to risk of harm to consumers, including reputational harm.[1662] This appears to be a fact-finding investigation, rather than an investigation of any

[1655] FTC, Comment Letter on Artificial Intelligence and Copyright to U.S. Copyright Office Docket No. 2023-6 (Oct. 30, 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/p241200_ftc_comment_to_copyright_office.pdf.

[1656] Press Release, FTC, FTC Launches Inquiry into Generative AI Investments and Partnerships (Jan. 25, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships; Dave Michaels, *FTC Launches Probe of Big Tech's AI Investments*, WALL ST. J. (Jan. 25, 2024), https://www.wsj.com/tech/ai/ftc-announces-ai-review-to-probe-roles-of-microsoft-open-ai-4255398a.

[1657] Dave Michaels & Tom Dotan, *FTC Opens Antitrust Probe of Microsoft AI Deal*, WALL ST. J. (Jun. 6, 2024), https://www.wsj.com/tech/ai/ftc-opens-antitrust-probe-of-microsoft-ai-deal-29b5169a.

[1658] Bedoya, supra note 1650, at 15–16.

[1659] *Id.*

[1660] *Id.*

[1661] Cat Zakrzewksi, *The FTC Investigates OpenAI Over Data Leak and ChatGPT's Inaccuracy*, WASH. POST (Jul. 23, 2023), https://www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/.

[1662] Government investigations are ordinarily confidential, but this was made public by the press. Civil Investigative Demand, FTC File No. 232-3044 (2023) at 2, https://www.washingtonpost.com/documents/67a7081c-c770-4f05-a39e-9d02117e50e8.pdf?itid=lk_inline_manual_4; *see also* Center for AI and Digital Policy, https://www.caidp.org/cases/openai/ (noting that the FTC sought information on bias, transparency, privacy, safety, and deception risk).

manifest violation, perhaps reflecting Chair Khan's desire to study and regulate technology when nascent, rather than waiting until it becomes mature.[1663]

c) Available remedies

The FTC, like other regulatory agencies, can pursue civil remedies in enforcement proceedings against companies. That includes penalties, cease and desist orders, and injunctive or other equitable relief, such as disgorgement, rescission, restitution, and corrective advertising.[1664]

"Algorithmic disgorgement" is a relatively novel form of disgorgement, one that may also be wielded against AI companies in cases involving the FTC Act's Section 5 violations for "unfair or deceptive acts or practices." "Algorithmic disgorgement," also referred to as "model deletion" or "model disgorgement," requires the offending company to give up improperly obtained data and the algorithm trained on such data (*see above section 4.1.3.D.*)[1665] The premise of algorithmic disgorgement mirrors that of any other regulatory disgorgement order: Companies who collect data illegally "should not be able to profit from either the data or any algorithm developed using it."[1666]

To date, the FTC has ordered model deletion of algorithms in five separate instances, beginning with the 2019 settlement with Cambridge Analytica.[1667] The most recent settlement, in December 2023, involved the destruction of an AI facial recognition model that Rite Aid pharmacy corporation used to surveil and identify customers who it believed were likely to engage in shoplifting or other problematic in-store behavior. The facial recognition model frequently misidentified individuals and generated thousands of false positive matches.[1668] Though incipient, algorithmic disgorgement offers a tailor-made remedy for consumer protection violations while also preventing future discrimination, loss of opportunity, and dignitary harms

---

[1663] Cecilia Kang and Kade Metz, *FTC Opens Investigation into ChatGPT Maker over Technology's Potential Harms*, N.Y. TIMES (Jul. 13, 2023), https://www.nytimes.com/2023/07/13/technology/chatgpt-investigation-ftc-openai.html.

[1664] *See* Appendix A, A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority, Federal Trade Commission (last updated May 2021), https://www.ftc.gov/about-ftc/mission/enforcement-authority.

[1665] Brandon Lalonde, *Explaining model disgorgement*, IAPP (Dec. 13, 2023), https://iapp.org/news/a/explaining-model-disgorgement/; *see also* Joshua A. Goland, *Algorithmic Disgorgement: Destruction of Artificial Intelligence Models as the FTC's Newest Enforcement Tool for Bad Data* (March 1, 2023). RICHMOND J. OF LAW AND TECH., Vol. XXIX, Issue 2 (2023), https://ssrn.com/abstract=4382254.

[1666] Rebecca Kelly Slaughter, *Algorithms and Economic Justice*, YALE J. L. & TECH. 37–39, https://yjolt.org/sites/default/files/23_yale_j.l._tech._special_issue_1.pdf.

[1667] Bruce D. Sokler, et al., *Algorithmic Disgorgement: An Increasingly Important Part of the FTC's Remedial Arsenal — AI: The Washington Report*, MINTZ (Jan. 24, 2024), https://www.mintz.com/insights-center/viewpoints/54731/2024-01-23-algorithmic-disgorgement-increasingly-important-pArticle.

[1668] Press Release, Rite Aid Banned from Using AI Facial Recognition After FTC Says Retailer Deployed Technology without Reasonable Safeguards, Federal Trade Commission (Dec. 19, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without.

that may be uniquely caused by AI and machine-learning models. Deleting both the data and the model would serve as a potent deterrent for abuses.[1669]

2) Consumer Financial Protection Bureau

Another regulatory agency that has been proactive in AI is the Consumer Financial Protection Bureau (CFPB), which was established to enforce federal consumer financial law and ensure fairness, transparency, and competition in the market for consumer financial products.[1670] A key area of concern for the CFPB has been the potential of AI (or predictive decision-making models more broadly) to make discriminatory or biased decisions in consumer financial matters, such as credit approval, mortgage lending, and home valuation. In April 2023, the CFPB, together with the FTC, the Equal Employment Opportunity Commission (EEOC), and the Civil Rights Division of the Department of Justice (DOJ), issued a joint statement pledging to use the agencies' existing authority to combat AI-based bias/discrimination in their respective regulatory domains. In remarks accompanying the joint statement, CFPB Director Rohit Chopra emphasized that "there is no exemption in our nation's civil rights laws for new technologies that engage in unlawful discrimination" and that companies cannot use a lack of understanding of their own algorithms as a defense against legal liability.[1671]

The CFPB has so far followed through on these pronouncements by promulgating a pair of policy circulars, one issued in May 2022 and the other in September 2023.[1672] In both, the agency established guidance to the industry on existing legal requirements and their applicability to the use of algorithmic or other predictive decision-making tools in consumer credit. The May 2022 circular emphasized that (a) federal consumer financial protection laws and notice/explanation requirements apply regardless of the technology used, and (b) the fact that an algorithm is complicated/new/opaque does not free companies from their obligation to comply with these requirements. The September 2023 circular added further specificity on these points. It stated that companies employing AI tools cannot merely provide broad or stock explanations to consumers that do not reflect the specific, accurate reasons for an adverse decision (e.g., denial or reduction of credit).

[1669] Jevan Hutson & Ben Winters, *America's Next "Stop Model!": Model Deletion*, 8 GEORGETOWN L. TECH. REV. 125, 126–28 (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4225003.

[1670] *About Us*, CFPB (last visited Apr. 28, 2023), https://www.consumerfinance.gov/about-us/.

[1671] Rohit Chopra, Director, CFPB, Prepared Remarks on the Interagency Enforcement Policy Statement on "Artificial Intelligence," (Apr. 25, 2023), https://www.consumerfinance.gov/about-us/newsroom/director-chopra-prepared-remarks-on-interagency-enforcement-policy-statement-artificial-intelligence/#2.

[1672] *Consumer Financial Protection Circular 2023-03*, CFPB (Sept. 19, 2023), https://www.consumerfinance.gov/compliance/circulars/circular-2023-03-adverse-action-notification-requirements-and-the-proper-use-of-the-cfpbs-sample-forms-provided-in-regulation-b/; Press Release, CFPB, CFPB Acts to Protect the Public from Black-Box Credit Models Using Complex Algorithms (May 26, 2022), https://www.consumerfinance.gov/about-us/newsroom/cfpb-acts-to-protect-the-public-from-black-box-credit-models-using-complex-algorithms/.

Similar themes appear in the CFPB's June 2023 white paper on "Chatbots in Consumer Finance."[1673] The report examines a broad spectrum of chatbots, ranging from rudimentary rule- or keyword-based systems to more advanced systems built atop LLMs. The report emphasizes that, "[i]n instances where financial institutions are relying on chatbots to provide people with certain information that is legally required to be accurate, being wrong may violate those legal obligations."[1674] And in a reference to the propensity of AI models to hallucinate plausible-sounding falsehoods, the report goes on to note that "the underlying statistical methods [of LLMs] are not well-positioned to distinguish between factually correct and incorrect data."[1675]

Accuracy requirements could present a significant compliance challenge for consumer finance companies, particularly those using large or complex AI models. The models' opacity even to the developers means that, in many instances, compliance may not be possible no matter how many technologists or lawyers they throw at the problem. This points to a more general lesson: Strong, legally enforceable requirements for accurate, specific explanations of algorithmic decisions—which may initially seem like relatively harmless disclosure exercises—could force companies to think twice about which algorithmic tools they deploy and in which contexts. If the penalties for noncompliance are substantial enough, e.g., algorithmic disgorgement, companies may refrain from deploying models in sensitive settings until they have a more advanced understanding of how the model makes decisions (and the state of algorithmic explainability more generally).

3) Other agencies

Other agencies have also indicated interest in taking AI-related actions. For example, as noted above, the FTC and CFPB were joined by the EEOC and DOJ Civil Rights Division in pledging to use their existing powers to combat AI-related bias and discrimination.[1676] The Equal Employment Opportunity Commission (EEOC) has also issued technical guidance on how AI's use in hiring and evaluation of employees could interact with requirements from two major anti-discrimination laws it enforces—the Americans with Disabilities Act[1677] and Title VII of the Civil Rights Act.[1678] The joint interagency statement and the EEOC's guidance are framed in terms of AI in general, rather than generative AI specifically. But they still apply to generative AI and could take on increasing importance as generative AI's capabilities expand the range of AI uses in employment/hiring contexts and the potential for discrimination and bias, if such deployment is not done with care.

[1673] *Chatbots in consumer finance* at 12, CFPB (Jun. 6, 2023), hthttps://files.consumerfinance.gov/f/documents/cfpb_chatbot-issue-spotlight_2023-06.pdf.

[1674] *Id.*

[1675] *Id.*

[1676] Rohit Chopra, et at., *Joint Statement on Enforcement Efforts against Discrimination and Bias in Automated Systems*, FEDERAL TRADE COMMISSION (Apr. 25, 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf.

[1677] *The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees*, EEOC (May 12, 2022), https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence.

[1678] *EEOC Releases New Resource on Artificial Intelligence and Title VII*, EEOC (May 18, 2023), https://www.eeoc.gov/newsroom/eeoc-releases-new-resource-artificial-intelligence-and-title-vii.

The Food and Drug Administration (FDA) is another US regulatory agency that has moved quickly regarding the use of AI in the industry it oversees. The FDA issued a paper in March 2024 on how the various public health agencies are collaborating to safeguard public health while fostering responsible and ethical innovation.[1679] As of May 2024, the FDA had already authorized 882 AI/ML-enabled medical devices,[1680] a category that includes both physical medical devices incorporating AI/ML software ("software *in* a medical device") and standalone medical software utilizing AI/ML ("software *as* a medical device").[1681] The FDA has also issued guidance that addresses a perhaps underappreciated challenge in AI governance – how regulators should deal with the fact that AI models are dynamic and evolve as they learn from new data.[1682] The FDA's recognition of the dynamic nature of AI models is potentially instructive, given calls from some quarters for a new regulatory agency to oversee an approval/licensing process that AI models must undergo before hitting the market.

5.3.2.B. The Biden Administration's strategy

Until the Biden administration promulgated the October 2023 Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,[1683] its strategy seemed geared toward encouraging the AI industry to adopt ethical practices – that is, to hortatory nudges rather than definitive law. Since then, the administration has taken a stronger, active hand not only in promoting dialogue with the private sector, but mobilizing the entire federal government to consider the potential of AI and take appropriate regulatory action.

1) The Blueprint for an AI Bill of Rights

In October 2022, a month before ChatGPT's release recalibrated the public's perceptions of AI's capabilities and risks, the White House's Office of Science and Technology Policy released a Blueprint for an AI Bill of Rights (the *Blueprint*).[1684] The *Blueprint* laid out five principles and associated practices that

[1679] *Artificial Intelligence & Medical Products: How CBER, CDER, CDRH, and OCP are Working Together*, FDA (March 2024), https://www.fda.gov/media/177030/download?attachment.

[1680] *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*, FDA (May 13, 2024), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

[1681] *Software as a Medical Device (SaMD)*, FDA (Dec. 4, 2018), https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd.

[1682] *CDRH Issues Draft Guidance on Predetermined Change Control Plans for Artificial Intelligence/Machine Learning-Enabled Medical Devices*, FDA (March 30, 2023), https://www.fda.gov/medical-devices/medical-devices-news-and-events/cdrh-issues-draft-guidance-predetermined-change-control-plans-artificial-intelligencemachine.

[1683] Exec. Order No. 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Fed. Reg. 75191, 75191–75226 (Nov. 1, 2023), https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[1684] Although ChatGPT was the first cutting-edge LLM to be publicly released, some image generators, such as OpenAI's own DALL·E 2, had been released earlier in 2022.

should guide the private sector's design, use, and deployment of AI "to protect the rights of the American public in the age of artificial intelligence."[1685]

1. safe and effective systems – "You should be protected from unsafe or ineffective systems."
2. algorithmic discrimination protections – "You should not face discrimination by algorithms, and systems should be used and designed in an equitable way."
3. data privacy – "You should be protected from abusive data practices via built-in protections, and you should have agency over how data about you is used."
4. notice and explanation – "You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you."
5. human alternatives, consideration, and fallback – "You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter."[1686]

This *Blueprint* has no legal force and does not create any new substantive or procedural legal rights. But its framing as a rights-based document was itself telling and foreshadowed the Biden administration's strategy toward AI. By choosing to announce its agenda with a consumer bill of rights, rather than a call for regulatory action directed to the rest of the government or a pro-innovation declaration to industry, the Biden administration signaled its awareness of the potential impact of AI on the public's rights, opportunities, and access to critical resources or services.[1687] Put another way, the *Blueprint* showed cognizance of the potential harms of this technology and a need for voluntary backstops led by industry, reinforced by the regulatory state. The *Blueprint* was measured about the risks of AI—not alarmist—and in proffering best practices for industry and regulation, it noted that all actions should be proportionate to the extent and nature of the harm or risk of harm.[1688] At the same time, the *Blueprint*'s rights-focused orientation overtly paid homage to antecedents like the OECD's 2019 Recommendation on AI (*see below section 6.2.1*)[1689] and, implicitly, to the European Commission's 2019 Ethics guidelines for trustworthy AI (*see above section 5.1.2*).[1690] So while the U.S. approach to AI would be uniquely American, it clearly drew inspiration from Europe.

And though structured as a consumer bill of rights and even phrased in the second person, as if the average reader is a consumer, the target audience was unmistakably the private sector. The clear objective was to nudge companies toward deploying AI systems responsibly through self-regulation, keeping in mind concerns about safety, discrimination/bias, privacy, transparency, and opt out rights.

2) Voluntary commitments

[1685] White House Office of Science and Technology Policy, Blueprint for an AI Bill of Rights (Oct. 2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.
[1686] *Id.* at 5–7.
[1687] *Id.* at 8.
[1688] *Id.*
[1689] *Id.* at 9.
[1690] High-Level Expert Group on Artificial Intelligence (EC), *Ethics Guidelines for Trustworthy AI*, EUROPEAN COMMISSION (Apr. 8, 2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

The Biden administration has prodded companies toward voluntary commitments and industry cooperation. Many companies had previously made commitments of their own accord, but the administration secured a uniform set of eight voluntary commitments from various leading AI companies, in two rounds. The first, in July 2023, involved seven companies: Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI.[1691] The second, in September 2023, involved eight companies: Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, and Stability.[1692] The commitments for the July batch of companies apply only to models more powerful than the then-industry frontier (GPT-4, Claude 2, PaLM 2, Titan, and DALL·E 2 for image generation).[1693] The September batch acceded to the commitments for any of their own future models more powerful than the most advanced model they had thus far produced.[1694] In other words, the commitments were forward-looking, not retroactive.

The eight commitments fall under three broad categories: safety, security, and trust.

- **Safety**. The companies committed to ensuring products are safe before introducing them to the public. As part of this, the companies pledged to (1) perform internal and external red teaming of models or systems to identify any potential for misuse, societal risks, and national security concerns. They also (2) agreed to share information among themselves and governments about trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards.

- **Security**. The second category was building systems that put security first. That entails (3) investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights. (All unreleased model weights are to be treated as core intellectual property for their business.) At the same time, companies were to (4) incentivize third-party discovery and reporting of issues and vulnerabilities (through mechanisms like bounty systems, contests, and prizes) to detect them even after internal red teaming.

- **Trust**. The third category was earning the public's trust through (5) developing and deploying mechanisms that enable users to understand if content is AI-generated, including provenance, watermarking, or both. Companies were also to (6) publicly report model capabilities, limitations, and domains of appropriate and inappropriate use. Looking ahead, the companies promised to (7)

[1691] Fact Sheet, White House, *Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage Risks Posed by AI* (Jul. 21, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

[1692] Fact Sheet, White House, Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage Risks Posed by AI (Sept. 12, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

[1693] White House, *Ensuring Safe, Secure, and Trustworthy AI* (July 2023), https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf.

[1694] White House, *Voluntary AI Commitments* (Sept. 2023), https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

prioritize research on societal risks posed by AI systems and (8) develop and deploy frontier AI systems to help address society's greatest challenges.[1695]

These voluntary commitments will remain effective until substantially similar regulatory measures addressing similar issues come into force.[1696] These voluntary measures are an important step indicative of high-level collaboration between government and industry, and industry acquiescence to the importance of innovating with safety, security, and trust in mind. Commentators have praised these efforts, even if critiquing them as vague and unambitious (i.e., they are things that leading AI companies are already doing).[1697] More generally, while voluntary commitments and a facilitated industry consortium are certainly valuable in calibrating and solidifying industry norms for responsible AI development, in the absence of legally binding obligations, they may be in tension with the competitive reality of companies trying to launch products as fast as possible.[1698] Even so, voluntary commitments still have some teeth. Even without a dedicated body to formally and legally examine companies' adherence to their commitments, NGOs, the media, and civil society groups do closely monitor company behavior and hold companies accountable in the public eye when they fail to meet their pledges. Moreover, the official and public adoption of certain commitments by these companies could potentially establish a standard of care or have other down-the-road legal consequences, so they still represent potent and meaningful policy advances.

3) Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

On October 30, 2023, the Biden administration released Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.[1699] President Biden's executive order directs or encourages action on AI from nearly every corner of the federal government.

Executive orders are directives issued by the president to the rest of the government and, therefore, cannot directly establish legally binding regulations, particularly law that governs the private sector.[1700] Nor can executive orders directly command or instruct the independent agencies that drive much of the federal regulatory apparatus, like the Federal Trade Commission (which is headed by a panel of commissioners whom the president appoints but cannot remove at will). As a result, the immediate specter of AI regulation following from the October 2023 Executive Order 14110 is rather limited. This reality was reflected in comments by both President Biden and key allies, such as U.S. Senate Majority Leader Chuck Schumer,

---

[1695] *Id.*

[1696] *Id.*

[1697] Kevin Roose, *How Do the White House's AI Commitments Stack Up?*, N.Y. TIMES (Jul. 22, 2023), https://www.nytimes.com/2023/07/22/technology/ai-regulation-white-house.html.

[1698] Nico Grant & Karen Wiese, *In A.I. Race, Microsoft and Google Choose Speed Over Caution*, N.Y. TIMES (Apr. 20, 2023), https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html.

[1699] Executive Office of the President [Joseph Biden]. Executive Order #14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Fed. Reg. 75191, 75191-75226 (Nov. 1, 2023), https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[1700] *See generally* Abigail A. Graber, *Executive Orders: An Introduction*, CONG. RESEARCH SERV. (Mar. 29, 2021), https://crsreports.congress.gov/product/pdf/R/R46738.

who emphasized at the time of issuance that the executive order would not be a substitute for Congress passing laws to regulate AI.[1701] Nevertheless, the executive order represents the most significant legal and policy action by the US federal government to date.

a) General overview

The length and breadth of the executive order reflects the ambitious scope of its eight guiding principles and policy priorities: (1) Ensuring the Safety and Security of AI Technology; (2) Promoting Innovation and Competition; (3) Supporting Workers; (4) Advancing Equity and Civil Rights; (5) Protecting Consumers, Patients, Passengers, and Students; (6) Protecting Privacy; (7) Advancing Federal Government Use of AI; and (8) Strengthening American Leadership Abroad.

The bulk of the executive order consists of instructions to various executive branch departments and agencies to develop and issue reports, guidelines, plans, and the like on issues related to these guiding principles in their respective domains. The executive order also sets out various mechanisms and bodies through which different entities within the federal government can coordinate with one another on AI issues, as well as solicit input from relevant stakeholders outside government. These efforts could form part of the factual or policy basis for future legally binding actions (e.g., by Congress or regulatory agencies). Stanford's Institute for Human-Centered Artificial Intelligence (HAI) has produced a tracker file[1702] of the full set of 150 requirements that agencies and other federal entities must implement. For brevity, this report discusses certain key provisions only, beginning with the definitions.

The Biden executive order includes a list of very precise, original definitions of technical terms (*see examples in Appendix VI*). A key definition relevant to understanding the obligations and policy emphases of the executive order is the term "dual-use foundation model." As defined in Section 3(k), this refers to a foundation model that exhibits "high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters." The executive order separately defines generative AI as "AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content."[1703]

Although the political focus on AI has been largely driven by the release of certain powerful generative AI tools (like ChatGPT or Stable Diffusion), much of the executive order is framed in terms of AI in general, rather than generative AI or foundation models in particular. However, some key sections of the executive

---

[1701] *See, e.g.*, John D. McKinnon et al., *Biden Taps Emergency Powers to Assert Oversight of AI Systems*, WALL ST. J. (Oct. 30, 2023), https://www.wsj.com/politics/policy/biden-to-use-emergency-powers-to-mitigate-ai-risks-cf7735d5; Cristiano Lima-Strong, *Schumer says 'only real answer' on AI is congressional action*, WASH. POST (Oct. 26, 2023), https://www.washingtonpost.com/technology/2023/10/26/schumer-artificial-intelligence-executive-order/.

[1702] Available at: https://docs.google.com/spreadsheets/d/1xOL4hkQ2pLR-IAs3awIiXjPLmhIeXyE5-giJ5nT-h1M/edit#gid=142633882. For a detailed discussion of the tracker, see Caroline Meinhardt et al., *By the Numbers: Tracking The AI Executive Order*, STANFORD HAI (Nov. 16, 2023), https://hai.stanford.edu/news/numbers-tracking-ai-executive-order.

[1703] The EO's general definition of "artificial intelligence" is the same one used in the National Artificial Intelligence Act of 2020, namely "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments." 15 U.S.C. 9401(3).

order—including those that create new reporting obligations for industry—are aimed specifically at powerful generative AI models and the infrastructure underlying them. That is, aside from sections on synthetic content, the most consequential sections of the executive order target "dual use foundation models," not merely "AI" or "generative AI."

b) Studies

Much of the executive order consists of the president directing various agencies of the federal government to examine the uses, benefits, and risks of AI in particular sectors/contexts, then issuing reports or other non-binding guidance on them. The executive order establishes (or directs federal departments to establish) a range of task forces and other bodies to conduct analysis and facilitate intragovernmental communication on AI-related topics, as well as mechanisms by which other stakeholders (e.g., industry, civil society) can offer input.

i) Study of the risks and governance strategies related to synthetic content

One of the many tasks the executive order assigns to the Secretary of Commerce is to study risks and governance strategies related to synthetic content. Specifically, the executive order calls for a report "identifying the existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques" for, among other things: authenticating and tracking content provenance, detecting and labeling synthetic content, and preventing generative AI from generating child sexual abuse material (CSAM) and nonconsensual intimate images.[1704] The Department of Commerce report is to be followed with guidance, updated periodically, on the state of such tools and how federal agencies may use them (e.g., for authenticating official government content).[1705]

ii) Study on the dual-source foundation models with publicly available model weights

Another notable feature of the October 2023 executive order is its treatment of powerful open-source models. Section 4.6 of the executive order observes that dual-use foundation models—models which can be used for both civilian and military purposes—with widely available model weights (i.e., "open source" models) can bring "substantial benefits to innovation, but also substantial security risks, such as the removal of safeguards within the model."[1706] It directs the Secretary of Commerce to examine the risks, benefits, and potential governance strategies (including "voluntary, regulatory, and international" approaches) for dual-use foundation models and to publish a report on its findings. The report is to be issued after soliciting input from stakeholders through a public consultation process, and the Biden administration's ultimate stance on open-source models will presumably be guided by the findings and recommendations of the report.[1707] The executive order and the language of the Commerce Department's February 2024 request for

[1704] EO 4.5(a).
[1705] EO 4.5(b)-(c).
[1706] EO 4.6.
[1707] Nat'l Telecommunications and Information Administration, Dep't of Commerce, 89 Fed. Red. 14059 (Feb. 26, 2024), https://www.federalregister.gov/documents/2024/02/26/2024-03763/dual-use-foundation-artificial-intelligence-models-with-widely-available-model-weights.

comment, which solicits input on nine key questions for open foundation models, suggest that there is a perception open models present distinctive risks and governance challenges, which may in turn require a distinct policy and regulatory response.

iii) Evaluation of CBRN [Chemical, Biological, Radiological, and Nuclear] threats

The executive order places significant emphasis on studying and mitigating AI-related threats to national security. It orders the Department of Homeland Security (DHS) to "evaluate the potential for AI to be misused to enable the development or production of CBRN [Chemical, Biological, Radiological, and Nuclear] threats" and to issue reports and studies based on this inquiry.[1708] The executive order highlights AI's potential role in the development of biological threats as an area of "particular focus,"[1709] perhaps motivated in part by warnings from leading AI figures, such as Dario Amodei, a co-founder of Anthropic.[1710]

c) Instructions to existing agencies

The executive order directs various departments of the executive branch to develop guidelines, standards, and best practices for AI safety and security.[1711] It also encourages independent agencies to consider (as they deem appropriate) their full range of existing authorities to address risks that may arise from the use of AI.[1712]

i) Instruction to the Federal Trade Commission (FTC)

Chief among the independent agencies charged with considering the risks of AI is the FTC. The executive order specifically mandated the FTC to evaluate whether to exercise its dual authority to ensure fair competition in the AI marketplace and to ensure that consumers and workers are protected from unfair and deceptive practices that may be enabled by the use of AI.[1713] The FTC's role was discussed earlier at Section 5.3.2.A.

ii) Instruction to the National Institute of Standards and Technology (NIST)

[1708] EO 4.4(a).

[1709] *Id.*

[1710] Dario Amodei has warned about the potential for foundation models to be used in the development of bioweapons. *See, e.g.,* Anna Edgerton and Oma Seddiq, *Anthropic's Amodei Warns US Senators of AI-Powered Weapons*, BLOOMBERG (Jul. 25, 2023), https://www.bloomberg.com/news/articles/2023-07-25/anthropic-s-amodei-warns-us-senators-of-ai-powered-bioweapons; Gerrit De Vynck, *AI leaders warn Congress that AI could be used to create bioweapons*, WASH. POST (July 25, 2023), https://www.washingtonpost.com/technology/2023/07/25/ai-bengio-anthropic-senate-hearing/.

[1711] EO § 4.1.

[1712] EO § 8.

[1713] EO § 5.3.

NIST, a non-regulatory agency under the United States Department of Commerce, is renowned for its expertise in developing standards for information security and cybersecurity.[1714] On January 26, 2023, NIST released the *AI Risk Management Framework (RMF) version 1.0*, using the same general template as NIST's prior *Cybersecurity Framework*[1715] and *Privacy Framework*.[1716] The AI RMF was designed to equip private and public AI actors with approaches that increase the trustworthiness of AI systems and help foster the responsible design, development, and use of AI systems over time.[1717] The RMF is voluntary, non-sector-specific, and use-case agnostic, which means organizations of all sizes and in all sectors can implement the approaches recommended by the Framework. The October 2023 Executive Order 14110 directs NIST to enhance this version 1.0 of its *AI RMF* by creating additional resources specifically for generative AI.[1718]

The core of the RMF are four specific functions which help organizations address the risks of AI systems in practice – govern, map, measure, and manage. Each one of those is broken down further into categories and subcategories.[1719] The RMF articulates the following seven characteristics of trustworthy AI systems: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.[1720] As its v1 appellation denotes, the RMF is an iterative, living document designed to be updated and supplemented by future guidance from NIST. The RMF and the accompanying Playbook provide detailed guidance on how organizations can assess and manage these risks in practice, organized by the govern, manage, map, and measure categories and subcategories. The Playbook in particular offers suggested actions and documentation for achieving the RMF's core outcomes.[1721]

President Biden's October 2023 Executive Order 14110 directed NIST to publish both the *AI RMF: Generative AI Profile* and the *Secure Software Development Framework (SSDF) for Generative AI and Dual-Use Foundation Models*.[1722] NIST introduced an initial draft of both on April 29, 2024 for public comment. The *AI RMF Generative AI Profile* addresses the risks associated with the specific use cases of generative AI. This document was shaped by the efforts of the Generative AI Public Working Group, which NIST established in July 2023.[1723] The *Generative AI Profile* aims to assist organizations in defining risks that are novel to or exacerbated by generative AI, categorizing them into 12 groups of specific risks ranging from chemical, biological, radiological, or nuclear (CBRN) information and other dangerous or abusive content to confabulations (hallucinations), data privacy, information integrity, and information security. The guidance concludes by providing a table of recommended actions to help organizations govern, map,

---

[1714] See *supra* note 32.

[1715] *Cybersecurity Framework*, NIST, https://www.nist.gov/cyberframework (last visited Apr. 28, 2024).

[1716] *Privacy Network*, NIST, https://www.nist.gov/privacy-framework (last visited Apr. 1, 2024).

[1717] NIST, ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK [hereinafter AI RMF] 2 (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[1718] EO § 4.1(a)(i)(A).

[1719] AI RMF at 2–3.

[1720] *Id.* at 12.

[1721] NIST, AI RMF PLAYBOOK (last visited June 1, 2024), https://airc.nist.gov/docs/AI_RMF_Playbook.pdf.

[1722] EO § 4.1(a)(i)(A) and 4.1(a)(i)(B).

[1723] This working group included over 2,500 members. NIST, NIST AI Public Working Groups, TRUSTWORTHY & RESPONSIBLE AI RESOURCE CENTER, https://airc.nist.gov/generative_ai_wg (last visited June 13, 2024).

measure, and manage these risks – corresponding to the RMF's core functions of governing, mapping, measuring, and managing.[1724] The *SSDF for Generative AI and Dual-Use Foundation Models* provides a common language for describing secure software development practices throughout the software development life cycle and augments the practices and tasks defined in SSDF version 1.1 by adding recommendations, considerations, notes, and informative references that are specific to generative AI and dual-use foundation model development.[1725]

Finally, NIST supplemented the executive order requests with two more related resources in late April 2024: *Reducing Risks Posed by Synthetic Content* and a *Plan for Global Engagement on AI Standards*. The first document surveys existing technical standards, tools, methods, and practices, as well as the potential development of future standards and techniques for authenticating content and tracking its provenance; labeling synthetic content (such as by watermarking); detecting synthetic content; and preventing generative AI from producing CSAM content or non-consensual deepfakes; testing software used for the aforementioned purposes; and auditing and maintaining synthetic content.[1726] The "Plan for Global Engagement on AI Standards" outlines U.S. policy efforts to coordinate with key international allies and partners on AI-related consensus standards, international cooperation and coordination, and information sharing.[1727] These two NIST documents were also open for public comment before they become final.[1728]

NIST guidance documents are not considered formal or informal regulatory action but have proven influential historically regardless. The *Cybersecurity Framework* is the best instantiation of this; it has become widely adopted by the private sector despite no formal legal action to make it so.[1729] So far, the *AI RMF* and associated documents have been just as influential. They have been incorporated by Microsoft,[1730]

[1724] ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK: GENERATIVE ARTIFICIAL INTELLIGENCE PROFILE, NIST AI 600-1 (Apr. 2024), https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf at 1-4.

[1725] NIST, SECURE SOFTWARE DEVELOPMENT PRACTICES FOR GENERATIVE AI AND DUAL-USE FOUNDATION MODELS, NIST SP 800-218A ipd (Apr. 2024), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.ipd.pdf at 1-2.

[1726] NIST, REDUCING RISKS POSED BY SYNTHETIC CONTENT, NIST AI 100-4 (Apr. 2024), https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf at 1.

[1727] NIST, A PLAN FOR GLOBAL ENGAGEMENT ON AI STANDARDS, NIST AI 100-5 (Apr. 2024), https://airc.nist.gov/docs/NIST.AI.100-5.Global-Plan.ipd.pdf at 2.

[1728] Press Release, Dep't of Commerce, Department of Commerce Announces New Actions to Implement President Biden's Executive Order on AI (Apr. 29, 2024), https://www.commerce.gov/news/press-releases/2024/04/department-commerce-announces-new-actions-implement-president-bidens.

[1729] Cameron F. Kerry, *NIST's AI Risk Management Framework plants a flag in the AI debate*, BROOKINGS INSTITUTE (Feb. 15, 2023), https://www.brookings.edu/articles/nists-ai-risk-management-framework-plants-a-flag-in-the-ai-debate/.

[1730] *Microsoft's AI Safety Policies: An update prepared for the UK AI Safety Summit*, MICROSOFT (Oct. 26, 2023), https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/#_edn54:~:text=and%20Technology%20(NIST)-,AI%20Risk%20Management%20Framework,-(RMF).%5B3.

IBM,[1731] and Anthropic,[1732] among many other companies. Just as many participated in the regulatory comment and feedback process. Some companies, such as Meta, Amazon, and Google DeepMind, have not openly announced that they have adopted the RMF or NIST guidance,[1733] but clearly have been influenced by it.[1734] Overall, judging by the feedback received during the comment periods, the sway of these documents in industry, and even a bipartisan Congressional proposal to make the RMF legally binding,[1735] NIST's voluntary guidance has been impactful by any measure.

iii) Instruction to the Department of Homeland Security

The executive order also tasks DHS with assessing the potential risks of AI to critical infrastructure (such as water supplies, power grids, telecommunications, and emergency management) and in cybersecurity, and incorporating the *AI RMF* and other security guidance into relevant safety and security guidelines for use by critical infrastructure owners and operators.[1736] DHS is also to establish an AI Safety and Security Board as an advisory committee comprising AI experts from the private sector, academia, and government.[1737] That DHS advisory board was constituted in April 2024 and includes the CEOs of OpenAI, Anthropic, Nvidia, Microsoft, and Alphabet, and other prominent chief executives and elected officials.[1738]

d) Reporting requirements on the basis of the Defense Production Act

The bulk of President Biden's executive order consists of instructions for federal departments to prepare non-binding reports, guidance, etc., as well as instructions that are binding only within the federal government itself. However, a few safety-related provisions of the executive order do establish binding

---

[1731] *IBM's Approach to Implementing the NIST AI RMF*, IBM (Sept. 26, 2023), https://www.ibm.com/policy/ibms-approach-to-implementing-the-nist-ai-rmf/.

[1732] Anthropic, *The Claude 3 Model Family: Opus, Sonnet, Haiku*, ANTHROPIC, https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf (Its approach to responsible release "draw[s] on guidance from the NIST *AI Risk Management Framework* and its Map, Measure, and Govern Subcategories.").

[1733] Anthony M. Barrett, Jessica Newman, & Brandie Nonnecke, *UC Berkeley AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models*, *Berkeley Center for Long-Term Cybersecurity* (Nov. 8, 2023), https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/.

[1734] Meta's Open Loop has a policy prototyping program in the United States which is focused on testing AI RMF 1.0. Generative AI Risk Management - Open Loop, OPENLOOP, https://www.usprogram.openloop.org/ (last visited June 16, 2024).

[1735] Press Release, Don Beyer, House of Representatives, Reps Lieu, Nunn, Beyer, Molinaro Introduce Bipartisan Bill To Establish AI Guidelines For Federal Agencies And Vendors (Jan. 10, 2024), https://beyer.house.gov/news/documentsingle.aspx?DocumentID=6066.

[1736] EO §§ 4.3(a)(i) and 4.3(a)(iii); *see also* Fact Sheet, Dep't of Homeland Sec., DHS Facilitates the Safe and Responsible Deployment and Use of Artificial Intelligence in Federal Government, Critical Infrastructure, and U.S. Economy (Apr. 29, 2024), https://www.dhs.gov/news/2024/04/29/fact-sheet-dhs-facilitates-safe-and-responsible-deployment-and-use-artificial.

[1737] EO § 4.3(a)(v).

[1738] Dustin Volz, *OpenAI's Sam Altman and Other Tech Leaders to Serve on AI Safety Board*, WALL ST. J. (Apr. 26, 2024), https://www.wsj.com/tech/ai/openais-sam-altman-and-other-tech-leaders-to-serve-on-ai-safety-board-7dc47b78?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=apple-openai-iphone-ai.

obligations for the private sector, in the form of reporting requirements for companies that develop powerful AI models or control the compute infrastructure on which such models depend.

Section 4.2(a) of the executive order references the Korean War-era Defense Production Act (1950), which gives the president broad authority to influence domestic industry for national security purposes.[1739] Under this authority, the executive order directs the Secretary of Commerce to establish two reporting requirements for private industry: one directed at developers of dual-use foundation models and the other at companies that control large compute clusters needed for the training of powerful models.

i) Reporting requirements for developers of dual-use foundation models

The reporting requirement for developers of dual-use foundation models applies only to models meeting a level of training compute just above what estimates suggest was used to train the current generation of frontier models (e.g., GPT-4); i.e., computing power greater than 10^{26} floating point operations per second (FLOPS).[1740] This threshold is slightly above the criterion used to determine general-purpose AI models "with systemic risk" under the AI Act, which is 10^{25} (*see section 5.1.2.C(2)*). With such a threshold, the reporting requirement appears intended to apply to the training/development of the *next* generation of frontier models (e.g., an eventual GPT-5) but not the *current* generation, such as GPT-4.[1741] That said, the compute threshold set by the executive order is a placeholder, with the Secretary of Commerce instructed to refine (and periodically update) the technical criteria that would trigger the reporting requirement.[1742]

Developers of dual-use foundation models meeting these criteria must provide the government with information on three major topics.

- First, they must report any ongoing or planned training/development of dual-use foundation models.
- Second, they must report the ownership and possession of model weights. For both training and weights, they must report cybersecurity and physical security measures used to protect the AI model and its weights from unauthorized access.
- Third, they must report the results of all red-teaming exercises and measures that the company has taken to improve the model's safety (including mitigations in response to vulnerabilities found during red teaming).[1743] Eventually, what counts as relevant red teaming subject to the reporting requirement will be based on the guidance for red teaming that NIST is instructed to develop elsewhere in the executive order.[1744]

ii) Reporting requirements for compute clusters

[1739] *See* Heidi M. Peters et al., *2022 Invocation of the Defense Production Act for Large-Capacity Batteries: In Brief*, CONG. RESEARCH SERV. (May 27, 2022), https://crsreports.congress.gov/product/pdf/R/R47124.

[1740] There is a lower threshold of 10^{23} FLOPS for models trained using primarily biological sequence data.

[1741] Rishi Bommasani et al., *Decoding the White House AI Executive Order's Achievements*, STANFORD HAI (Nov. 2, 2023), https://hai.stanford.edu/news/decoding-white-house-ai-executive-orders-achievements (estimating that GPT-4 is just shy of this threshold).

[1742] EO 4.2(b).

[1743] EO 4.2(a)(i)(A)-(C).

[1744] Specifically, section 4.1(a)(ii).

The other reporting requirement established under the authority of the Defense Production Act deals with compute clusters. It requires companies or other entities that acquire, develop, or possess large-scale computing clusters to report this to the government, including the location and total computing power of each cluster.

FIGURE 46. Reporting requirements on statutory basis from Executive Order 14110[1745]

Provision	Implementing department/agency (primary)	Covered entities	Scope or threshold/trigger	Contents
4.2(a)(i)	Commerce	Companies developing or demonstrating an intent to develop potential dual use foundation models (DUFMs)	Initial: any model trained using computing power greater than 10^{26} floating point operations per second (FLOPS) - lower threshold of 10^{23} FLOPS for models trained using primarily biological sequence data Eventual: technical thresholds for models to be developed (and regularly updated) by Department of Commerce	Companies must report: - ongoing or planned training/development of DUFMs - ownership and possession of model weights - measures taken to ensure integrity of the training process against physical and cybersecurity threats - results of any red-team testing (applying guidance to be developed by NIST pursuant to EO) - any measures the company has taken to meet safety objectives (such as mitigations to improve red-teaming performance)
4.2(a)(ii)	Commerce	Companies, individuals, organizations, or other entities that acquire, develop, or	Initial: any computing cluster with theoretical maximum computing capacity of 10^{20}	Covered entities must report: - acquisition, development, or possession of any such clusters

[1745] For further details, see White House, Fact Sheet, President Biden Issues Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Oct. 30, 2023), and the executive order (EO) itself, Exec. Order No. 14110, *supra* note 1527.

		possess a potential large-scale computing cluster	FLOPS for AI training, located in a datacenter with network connectivity >100 Gbit/s <u>Eventual:</u> technical thresholds for computing clusters to be developed (and regularly updated) by Department of Commerce	- existence and location of the clusters - amount of total computing power in each cluster

e) Other reporting requirements; US "infrastructure as service" providers

President Biden's executive order contains another set of reporting requirements that applies to US providers of "infrastructure as a service" (IaaS, i.e., cloud computing companies). The legal basis for using presidential authority to impose such reporting requirements is "the national emergency related to significant malicious cyber-enabled activities," declared in previous executive orders during the administration of President Barack Obama.[1746]

Specifically, President Biden's October 2023 executive order directed the Secretary of Commerce to propose regulations that would require US providers of IaaS to report any transactions where foreign persons or entities use the infrastructure of IaaS companies to train "large AI model[s]" that could be used for malicious cyber-enabled activity.[1747] Under those regulations, which were proposed by Commerce in January 2024, an IaaS product is any product or service that provides processing, storage, networks, or other fundamental computing resources with which a consumer could deploy and run software that is not predefined for a specific purpose.[1748]

The proposed regulations explicitly state they would cover both managed products and services (where the provider is responsible for aspects of system configuration or maintenance) and unmanaged ones (where the provider is responsible only for ensuring that the product is available). The proposed regulations also cover virtualized products and services (in which computing resources are split between virtualized computers) and dedicated ones (in which the total computing resources of a physical machine are provided

[1746] EO 4.2(c).

[1747] EO 4.2(c)(i).

[1748] Taking Additional Steps to Address the National Emergency with Respect to Significant Malicious Cyber-Enabled Activities, 89 Fed. Reg. 5698, 5726 (proposed Jan. 29, 2024) (to be codified at 15 C.F.R. pt. 7).

to a single person). As a result, the proposed regulations would likely span all cloud companies identified in Figure 4 (*see above section 2.3.1.*) as infrastructure providers.[1749] IaaS providers would be obliged to develop and maintain a written customer identification program that establishes procedures for identifying and verifying foreign user accounts. The proposed regulations are in the comment period and await finalization later in 2024.

FIGURE 47. Requirements for IaaS providers in the Executive Order 14110[1750]

EO provision	Implementing department/agency (primary)	Covered entities	Scope or threshold/trigger	Contents
4.2(c)	Commerce	US IaaS companies	Foreign persons transacting with a US IaaS provider to train a large AI model with potential capabilities that could be used in malicious cyber-enabled activity - Department of Commerce instructed to develop technical criteria for when a model "could be used in malicious cyber-enabled activity"	Department of Commerce instructed to propose regulations requiring US IaaS providers to: - submit a report to Commerce whenever such a transaction occurs (report must include the identity of the foreign person(s), existence of the training run, and other information as determined by Commerce) - prohibit any foreign resellers of their IaaS products from offering the products unless they also comply with the reporting requirements.

f) Evaluating the executive order's present and future impact

Beyond its mobilization of the government, the 100-plus page executive order demonstrates how seriously the Biden administration takes its responsibility to foster a vibrant AI ecosystem in the United States, while also drawing parameters to harness and govern that ecosystem. In the estimation of scholars at the Stanford HAI, it is a "major step forward to ensure that America remains at the forefront of responsible

[1749] *Id.* at 5702.

[1750] *See* Exec. Order No. 14110, *supra* note 1527.

innovation."[1751] For an executive order, it managed to be remarkably precise, naming over 50 federal entities and producing around 150 distinct action items for them. At the same time, it is also broad – covering nearly every known potential aspect of this nascent technology and mobilizing the full government to modernize and respond holistically to it.[1752] And 90 days after the executive order's issuance, the executive branch had made significant progress in completing nearly all of the action items due.[1753] Moreover, according to Stanford University's Human-Center AI (HAI), the administration has been "admirably transparent" about its progress.[1754]

On substance, the executive order has been hailed for its middle-of-the-road approach, for instance, by favoring openness and eschewing a licensing regime.[1755] On the other hand, critics charge that the executive order has a "glaring absence" of transparency requirements around model development.[1756] Nor does it take the potential harms of open-source models seriously enough.[1757] These critiques perhaps hold the executive branch to an exacting standard; after all, executive orders are, by nature, limited in scope and effect. And the next president could rescind this one entirely and start from scratch. In any event, the executive order does not (and cannot) directly or comprehensively regulate the private sector. Nevertheless, the executive order is an important advance and takes several first steps that lay the foundation for future regulatory structures. Its ultimate success or failure will be judged by the extent to which it galvanizes regulatory agencies to begin constructing those structures and whether that, in turn, requires Congress to enact fresh legislation.

As of this writing, the impact of President Biden's executive order is beginning to reverberate throughout the federal government. The executive order has three thematic aims: strengthening AI governance, advancing responsible AI innovation, and managing risks from the use of AI. The administration expanded those goals to also include expanding transparency of AI use and growing the AI workforce.[1758] To further implement these objectives, in March 2024, the White House's Office of Management and Budget (OMB) ordered every federal agency to, among other things: (i) designate a chief AI officer, (ii) convene an internal

---

[1751] Rishi Bommasani et al., *Decoding the White House AI Executive Order's Achievements*, see *supra* note 1741.

[1752] *Id.*

[1753] Caroline Meinhardt et al., *Transparency of AI EO Implementation: An Assessment 90 Days In,* STANFORD HAI (Feb. 21, 2024), https://hai.stanford.edu/news/transparency-ai-eo-implementation-assessment-90-days.

[1754] *Id.*

[1755] Arvind Narayanan, Sayash Kapoor & Rishi Bommasani, *What the executive order means for openness in AI*, (Oct. 31, 2023), https://www.aisnakeoil.com/p/what-the-executive-order-means-for.

[1756] *Id.*

[1757] Renée Diresta & Dave Willner, *White House AI Executive Order Takes on Complexity of Content Integrity Issues*, TECH POLICY.PRESS (Nov. 1, 2023), https://www.techpolicy.press/white-house-ai-executive-order-takes-on-complexity-of-content-integrity-issues/; Casey Newton, *Biden Seeks to Rein in AI*, PLATFORMER (Oct. 31, 2023), https://www.platformer.news/biden-seeks-to-rein-in-ai/.

[1758] Fact Sheet, White House, Vice President Harris Announces OMB Policy to Advance Governance, Innovation, and Risk Management in Federal Agencies' Use of Artificial Intelligence (Mar. 28, 2024), https://www.whitehouse.gov/briefing-room/statements-releases/2024/03/28/fact-sheet-vice-president-harris-announces-omb-policy-to-advance-governance-innovation-and-risk-management-in-federal-agencies-use-of-artificial-intelligence/.

AI governance body, (iii) publicly release a strategy for the agency to identify and remove barriers to the responsible use of AI within the agency in the next year, and (iv) to implement certain minimum practices to avoid risks from "safety-impacting" and "rights-impacting" AI (*see Appendix IX*).[1759] With respect to (iv), the OMB memorandum defines safety-impacting AI as that which can significantly impact human life or well-being, the climate or environment, critical infrastructure, or strategic assets and resources. It also defines rights-impacting AI as that which has a significant effect on civil rights or liberties, equal opportunities, or access to critical or government resources.[1760] Federal agencies are directed to implement the minimum practices for safety-impacting and rights-impacting AI before December 1, 2024. Those practices include completing an AI impact assessment, testing the AI for performance in a real-world context, ongoing monitoring, mitigation of risks to rights and safety, and public notice and plain-language documentation.[1761]

Finally, the Biden administration has created an AI Safety Institute (AISI) under the aegis of NIST to facilitate government-industry cooperation on AI safety. The AISI's goals are to: advance the science of AI safety; articulate, demonstrate, and disseminate the practices of AI safety; and support institutions, communities, and coordination around AI safety. The AISI intends to work closely with diverse AI industry, civil society members, and international partners to achieve these goals.[1762] One of AISI's signature initiatives is the creation of the U.S. AI Safety Institute Consortium (AISIC)[1763] to "unite AI creators and users, academics, government and industry researchers, and civil society organizations in support of the development and deployment of safe and trustworthy artificial intelligence (AI)."[1764] AISIC will develop guidelines for red teaming, capability evaluations, risk management, safety and security, and watermarking synthetic content, among other things. AISIC counts more than 200 inaugural members, including OpenAI, Google, Anthropic, Amazon.com, Microsoft, Meta, Nvidia, Palantir, Intel, JPMorgan Chase, and Bank of America. The creation of the consortium hardens the commitments of industry participants and ensures that there will be a standing forum for discussion and setting standards.[1765]

5.3.2.C. Proposals for future legislation

[1759] Office of Mgmt. & Budget, Exec. Office of the President, M-24-10, Mem. For the Heads of Executive Departments and Agencies (Mar. 28, 2024), https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf.

[1760] *Id.* at 29–30. Certain purposes are presumed to be safety-impacting and rights-impacting. *Id.* at 31-33.

[1761] *Id.* at 17–23.

[1762] *Strategic Vision*, U.S. ARTIFICIAL INTELLIGENCE SAFETY INSTITUTE AT NIST, https://www.nist.gov/aisi/strategic-vision (last visited May 21, 2024).

[1763] *Artificial Intelligence Safety Institute Consortium*, NIST (Apr. 15, 2024), https://www.nist.gov/artificial-intelligence-safety-institute/artificial-intelligence-safety-institute-consortium-aisic.

[1764] Press Release, Dept. of Commerce, Biden-Harris Administration Announces First-Ever Consortium Dedicated to AI Safety (Feb. 8, 2024), https://www.commerce.gov/news/press-releases/2024/02/biden-harris-administration-announces-first-ever-consortium-dedicated.

[1765] *Id.* The consortium also comprises civil society participants, state and local governments, and academic teams, like Stanford University's Institute for Human-Centered AI, its Center for Research on Foundation Models, and its Regulation, Evaluation, and Governance Lab.

As of now, the proposals for future federal legislation fall into two types: broad frameworks for comprehensive regulation with identified policy priorities and desiderata, and more narrow draft legislation designed to cure known ills that have already manifested with generative AI.

1) Broad frameworks

The earliest two frameworks that gained sway in 2023 were the SAFE[1766] Innovation Framework from Senate Majority Leader Chuck Schumer (and its successor, the Innovation Roadmap), and the bipartisan framework on AI legislation from Senators Richard Blumental and Josh Hawley. Those frameworks set the stage for future legislation but have receded in significance in 2024 as two newer contenders have emerged in their place.

a) SAFE Innovation Framework and Innovation Roadmap

The "SAFE Innovation Framework" began its life in June 2023 as a one-page preview of Senator Schumer's policy goals for future AI legislation. The framework contains no concrete proposals or draft language readily translatable into law.[1767] Instead, it set forth five key policy objectives that should guide the US's approach to safe development of generative AI: (1) security (principally national security), (2) accountability, (3) foundations (model alignment), (4) explain (model transparency), and (5) innovation.[1768] At one page in length, it was light on details, but Senator Schumer used it as a springboard to convene a series of "AI Insight Forums" intended to educate lawmakers about artificial intelligence. These closed-door forums connected legislators and key AI experts to help shape the direction of future legislation.[1769] Such proactive efforts appeared to be an attempt to combat perceptions that Congress lacked an understanding of Web 2.0, contributing to its inability to properly formulate a policy response before it was too late.[1770]

The first AI Insight Forum was held in September 2023 and was successful in drawing nearly all major tech CEOs to Washington with considerable fanfare—though without the pressure of a formal hearing—to candidly discuss the future of AI, its opportunities, and possible regulation. The forum was also attended by labor union leaders and civil society representatives, who sat side by side with technology's captains of industry. Though there were disagreements among them—about which risks were most salient and how

[1766] An acronym for the first four principles of the framework: Security, Accountability, Foundations, and Explainability.

[1767] Karoun Demirjian, *Schumer Lays Out Process to Tackle A.I., Without Endorsing Specific Plans*, N.Y. TIMES (Jun. 21, 2023), https://www.nytimes.com/2023/06/21/us/ai-regulation-schumer-congress.html.

[1768] Charles Schumer, U.S. Senate, SAFE Innovation Framework (Jun. 21, 2023), https://www.democrats.senate.gov/imo/media/doc/schumer_ai_framework.pdf; *see also id.*

[1769] Gabby Miller, *US Senate AI 'Insight Forum' Tracker* (Dec. 8, 2023), TECH POLICY PRESS, https://www.techpolicy.press/us-senate-ai-insight-forum-tracker/.

[1770] *See, e.g.*, Emily Stewart, *Lawmakers seem confused about what Facebook does – and how to fix it*, VOX (Apr. 10, 2018), https://www.vox.com/policy-and-politics/2018/4/10/17222062/mark-zuckerberg-testimony-graham-facebook-regulations.

open-source models should be treated—there was "striking unanimity" on the need for American leadership on AI and broad agreement on the need for regulation of some sort.[1771]

After holding this first AI Insight Forum, Senator Schumer banded together with three other senators from both sides of the aisle to form a bipartisan Senate AI Working Group and hosted eight more forums. This Senate AI Working Group promulgated an innovation roadmap in May 2024 as a successor document to the SAFE Innovation Framework. The innovation roadmap was a redoubled attempt at SAFE Innovation and encouraged the government to promote AI innovation through various means, including future annual appropriations of at least $32 billion allocated across several federal agencies. It articulated broad policy goals of, among other things: investing in AI research and development, retraining the workforce, developing and enforcing new AI laws and guidelines which would preserve the integrity of elections, safeguarding against other AI-related risks, and ensuring cybersecurity and national security are protected.[1772] Senator Schumer and the AI Working Group have indicated that their plan is to have Congressional committees move smaller piecemeal legislation forward rather than wait for a larger package of AI-only legislation to come together.[1773] Thus, though the innovation roadmap presents an expanded vision with more specifics, it—much like its predecessor—does not illuminate a pathway for converting platitudes of "getting [AI regulation] right"[1774] into workable law.

b) Blumenthal-Hawley "Bipartisan framework for U.S. AI Act"

Senators Richard Blumenthal and Josh Hawley (who both sit on the Senate Judiciary Subcommittee on Privacy, Technology, and the Law) held a series of formal Congressional hearings on AI in September 2023 and proposed their own framework for AI legislation. Their *Bipartisan Framework for U.S. AI Act* contains more specifics than SAFE Innovation and focuses on preventing harms and ensuring accountability from AI companies through a combination of licensing, legal accountability, and transparency mechanisms. The five main planks of the Blumenthal-Hawley framework are to:[1775]

1. **"Establish a Licensing Regime Administered by an Independent Oversight Body.**" Registration with the oversight body would be required both for "sophisticated general purpose AI models" (e.g., GPT-4) and "models used in high-risk situations."[1776] The proposal of a

[1771] *See, e.g.*, Cecilia Kang, *In Show of Force, Silicon Valley Titans Pledge 'Getting This Right' With A.I.*, N.Y. TIMES (Sept. 13, 2023), https://www.nytimes.com/2023/09/13/technology/silicon-valley-ai-washington-schumer.html.
[1772] Bipartisan Senate AI Working Group, Driving U.S. Innovation In Artificial Intelligence (May 2024), https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf.
[1773] Ursula Perano, *Bipartisan group of senators unveil long-awaited guidance on AI bills*, POLITICO (May 15, 2024), https://www.politico.com/live-updates/2024/05/15/congress/schumers-roadmap-on-ai-bills-00157828.
[1774] Cecilia Kang, *In Show of Force, Silicon Valley Titans Pledge 'Getting This Right' With A.I.*, N.Y. TIMES (Sept. 13, 2023), https://www.nytimes.com/2023/09/13/technology/silicon-valley-ai-washington-schumer.html (quoting OpenAI's Sam Altman).
[1775] Richard Blumenthal, U.S. Senate, Bipartisan Framework for U.S. AI Act (Sept. 7, 2023), https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf.
[1776] *Id.*

dedicated licensing body echoes suggestions from figures such as OpenAI's Sam Altman.[1777] And imposing the registration requirement on both "general purpose" models and "high-risk" use cases parallels the EU's approach under the AI Act. However, both the creation of the oversight body and the scope of the registration requirement would likely receive pushback from politicians and business interests alike against overly restrictive regulation and its effects on innovation. Scholars have also raised concerns about the effectiveness and viability of disclosures, registration, licensing, and auditing proposals on technical and institutional feasibility grounds.[1778]

2. **"Ensure Legal Accountability for Harms."** The framework urges that "Congress should require AI companies to be held liable through entity enforcement and private rights of action when their models and systems" cause various harms.[1779] This element of the framework currently has draft legislation accompanying it.[1780] The senators had already introduced a bill in June 2023 that would amend Section 230 of the Communications Decency Act to clarify that its liability shield does not apply to AI-generated content.[1781]

3. **"Defend National Security and International Competition."** This element of the framework would see Congress using export controls, sanctions, and other tools to prevent the transfer of advanced AI models and equipment to countries that are US adversaries (China and Russia are specifically named) or major human rights violators. The proposal here dovetails with measures that the Biden administration is already pursuing through executive branch action, particularly export controls and investment restrictions aimed at limiting China's AI capabilities.[1782]

4. **"Promote Transparency."** The framework proposes a transparency regime that would require AI developers to "disclose essential information about training data, limitations, accuracy, and safety of AI models to users and other companies." It would also require notification of users when they are interacting with an AI system, as well as creation of a public database that is maintained by the new oversight body, to report adverse incidents and harms.

[1777] Cecilia Kang, *OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing*, N.Y. TIMES (May 16, 2023), https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

[1778] *See* Neel Guha et al., *AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing*, GEORGE WASHINGTON L. REV. (forthcoming), https://dho.stanford.edu/wp-content/uploads/AI_Regulation.pdf.

[1779] *Id.*

[1780] A bill to waive immunity under section 230 of the Communications Act of 1934 for claims and charges related to generative artificial intelligence, S.1993, 118th Cong. (2023).

[1781] Press Release, Josh Hawley, U.S. Senate, Hawley, Blumenthal Introduce Bipartisan Legislation to Protect Consumers and Deny AI Companies Section 230 Immunity (June 14, 2023), https://www.hawley.senate.gov/hawley-blumenthal-introduce-bipartisan-legislation-protect-consumers-and-deny-ai-companies-section.

[1782] Peter Baker and David E. Sanger, *Biden Orders Ban on New Investments in China's Sensitive High-Tech Industries*, N.Y. TIMES (Aug. 9, 2023), https://www.nytimes.com/2023/08/09/us/politics/biden-ban-china-investment.html.

5. **"Protect Consumers and Kids."** This element of the framework states that consumers "should have control over how their personal data is used in AI systems." It proposes "strict limits" on generative AI products/outputs involving kids, seeks implementation of "safety brakes" for high-risk applications, and notice to users when AI is being used to make adverse decisions.

In many respects, the Blumenthal-Hawley framework is the opposite of the "SAFE Innovation Framework." It is highly ambitious, suggesting a licensing regime under the purview of a new independent oversight body; demarcates rules of civil liability; and clarifies that Section 230's liability shield does not protect AI models. It would also impose a transparency obligation for AI model developers and deployers. In the words of one AI safety-focused policy organization, it is the spark for comprehensive AI regulation "that America desperately needs."[1783] Perhaps because of its sweeping nature, the Blumenthal-Hawley framework has failed to gain traction or attract support from other legislators who may favor a lighter touch.

c) Framework for mitigating extreme AI risks

In April 2024, a bipartisan group of four more senators, headlined by Senator Mitt Romney, released a more modest proposal focused squarely on addressing the most extreme risks of AI while avoiding full-blown comprehensive regulation, as the Blumenthal-Hawley bill envisioned. The Romney group's *Framework for Mitigating Extreme AI Risks* prioritizes the national security implications of AI while preserving innovation and the domestic AI industry.[1784] The framework would require:

1. entities that sell or rent hardware to report large acquisitions or usage of computing resources, particularly by foreign persons, to an oversight entity;
2. developers to notify the oversight entity when developing a frontier model and to incorporate safeguards against biological, chemical, cyber, and nuclear risks; and
3. developers to obtain a license from the oversight entity before deploying a model, to certify that the model has sufficient safeguards against those four risks.

The licensing could be tiered such that low-risk models could be made available on an open-source basis, whereas the highest-risk models could be deployed only to vetted customers. Interestingly, the framework left open who might carry out the oversight function –it could be a new interagency coordinating body, a preexisting federal agency, or a new agency altogether.[1785] In essence, this framework is a lighter version of Blumenthal-Hawley.

The preceding three frameworks were all from the Senate. A fourth framework may emerge from the U.S. House. House Speaker Mike Johnson and Minority Leader Hakeem Jeffries formed a bipartisan task force

---

[1783] *Strengths of Hawley and Blumenthal's Framework*, CENTER FOR AI POLICY (Oct. 24, 2023), https://www.aipolicy.us/work/strengths-of-hawley-and-blumenthals-framework.

[1784] Press Release, Mitt Romney, U.S. Senate, Romney, Reed, Moran, King Unveil Framework to Mitigate Extreme AI Risks, (Apr. 16, 2024), https://www.romney.senate.gov/romney-reed-moran-king-unveil-framework-to-mitigate-extreme-ai-risks/.

[1785] Mitt Romney, U.S. Senate, Framework for Mitigating Extreme Risks (Apr. 16, 2024), https://www.romney.senate.gov/wp-content/uploads/2024/04/AI-Framework_2pager.pdf.

in February 2024 to produce a comprehensive report to expound high-level principles, forward-looking recommendations, and policy proposals.[1786]

Reconciling the competing frameworks from different camps in the Senate with whatever emerges from the House task force may be a tall order, as there appears to be a considerable chasm on policy objectives and priorities among different groups on Capitol Hill. There is no settled consensus on what kind of oversight body is called for or whether a licensing regime is necessary. And there is no agreement on other substantive legal issues, like model transparency, copyright, liability, and data protection. Consequently, more targeted narrow legislation that seeks to cure specific and manifest harms of AI bears more promise of adoption in the near to medium term.

2) Targeted legislation

According to an April 2024 count taken by the Brennan Center for Justice, 20 different pieces of AI-related legislation have been proposed during the current (118th) Congress.[1787] Many of the bills seek to direct federal agencies to take certain actions under existing authority or simply establish new standards under the same. Few are, or purport to be, comprehensive AI regulation, but most of the legislation would not dramatically reshape the regulatory environment for AI model development by creating a licensing regime or banning the training of models on copyrighted information. In any case, none appears particularly close to passage. Nevertheless, the introduced bills provide a sense of the concerns percolating among US policymakers today. And concepts introduced in the bills, if not wholesale provisions, may ultimately be carried forward and reflected in final legislation at some future time. Below is a summary of some other of the leading proposals:

a) Algorithmic Accountability Act of 2023

Originally introduced in 2019, this bill was reintroduced in September 2023 by Senator Ron Wyden and repurposed for AI as "The Algorithmic Accountability Act." This bill eschews overreach and, instead, targets any automated decision-making that might affect critical decisions about Americans' health, finances, employment, housing, and educational opportunities. The Algorithmic Accountability Act would require covered entity deployers—whether or not they are also developers—of automated decision systems to conduct initial impact assessments of their use of such systems. It would require the deployers to submit ongoing annual summaries of the impact assessments to the FTC, which would also be empowered to conduct rulemakings to enforce the regulations it would promulgate. This bill would also add 75 staff to the

[1786] Press Release, Hakeem Jeffries, House of Representatives, House Launches Bipartisan Task Force on Artificial Intelligence (Feb. 20, 2024), https://democraticleader.house.gov/media/press-releases/house-launches-bipartisan-task-force-artificial-intelligence.

[1787] Artificial Intelligence Legislation Tracker, BRENNAN CENTER (last updated Apr. 1, 2024), https://www.brennancenter.org/our-work/research-reports/artificial-intelligence-legislation-tracker. A press source suggests more than 170 Congressional bills in the last year alone have mentioned AI. *See* Brian Fung, *AI could disrupt the election. Congress is running out of time to respond*, CNN (Feb. 14, 2024), https://www.cnn.com/2024/02/14/tech/ai-bill-us-presidential-election/index.html.

FTC under a new Bureau of Technology charged with implementing the law. The bill expressly does not preempt state law.[1788]

b) AI Research, Innovation, and Accountability Act of 2023

In November 2023, a bipartisan group led by Senators John Thune and Amy Klobuchar introduced the AI Research, Innovation, and Accountability Act of 2023. This bill would instruct NIST to carry out research to facilitate provenance standards that allow users to distinguish between human-generated and AI-generated content. It also would direct NIST to support standardization for detecting and understanding emergent properties in AI systems, in order to mitigate issues from the unanticipated behavior of foundation models. This Accountability Act proposes new definitions for AI systems and propounds a distinction between developers of AI systems and deployers, imposing greater obligations on deployers. Finally, the bill would create transparency requirements and a certification framework for "critical-impact AI systems" in which organizations deploying such systems would submit annual reports on the design and safety of AI models to the Commerce Department and self-certify compliance with standards prescribed by Commerce.[1789]

c) AI Foundation Model Transparency Act

Representatives Anna Eshoo and Donald Beyer, who serve as co-chair and vice-chair of the House Congressional AI Caucus, respectively, introduced the AI Foundation Model Transparency Act in December 2023. The Act would direct the FTC to promulgate transparency standards for AI model deployers about training data and algorithms used in their models. Deployers of foundation models would have to produce disclosures to the public and the FTC.[1790] They would be required to report the sources of their training data, how the data are retained during the inference process, and describe the limitations or risks of the model. They would also have to explain how the model aligns with NIST's *AI RMF*, provide information on the computational power used to train and run the model, and report on efforts to red team the model to prevent it from providing inaccurate or harmful information.

Apart from demystifying the inner workings of the model and helping users understand their results, limitations, and potential biases, the Foundation Model Transparency Act aims to help copyright owners

---

[1788] Algorithmic Accountability Act of 2023, H.R. 2892, 118th Cong. (2023), https://www.wyden.senate.gov/imo/media/doc/algorithmic_accountability_act_of_2023_summary.pdf; *Lawmakers Reintroduce Bill to Regulate Use of AI Systems*, GOVERNMENT TECHNOLOGY (Sept. 25, 2023), https://www.govtech.com/artificial-intelligence/lawmakers-reintroduce-bill-to-regulate-use-of-ai-systems.

[1789] Artificial Intelligence Research, Innovation, and Accountability Act of 2023, S. 3312, 118th Cong. (2023); John Thune, U.S. Senate, One-pager summarizing the Artificial Intelligence (AI) Research, Innovation, and Accountability Act of 2023 (Nov. 15, 2023), https://www.thune.senate.gov/public/_cache/files/8c63ff8a-f528-4214-84f6-61bfa5665cec/9CE3283C53BE64087BBCF506E10C8FE3.artificial-intelligence.pdf.

[1790] Press Release, Don Beyer, House of Representatives, Beyer, Eshoo Introduce Landmark AI Regulation Bill (Dec. 22, 2023), https://beyer.house.gov/news/documentsingle.aspx?DocumentID=6052; Don Beyer, One-pager summarizing the AI Foundation Model Transparency Act (Dec. 22, 2023), https://beyer.house.gov/uploadedfiles/one-pager_ai_foundation_model_transparency_act_.pdf.

assess whether their rights have been infringed upon by the training of foundation models.[1791] If adopted, this bill would create 20 distinct transparency requirements for deployers, surpassing the five called for in the White House Executive Order (which some criticized as a paltry number). The 20 requirements in the AI Foundation Model Transparency Act are still fewer than the EU AI Act's 30, including 12 that this proposal did not capture.[1792]

d) Legislation targeting AI deepfakes and nonconsensual use of digital images

Several proposed AI bills target the rising proliferation of AI deepfakes, with more bills surely in the offing.

Two bills entitled AI Labeling Act of 2023 address concerns about deepfake images.[1793] The bills seek to require AI model developers to: include clear and conspicuous disclosures identifying AI-generated content, place metadata in AI outputs marking it as AI-generated, and to take reasonable steps to ensure downstream licensees of models do the same.

Other bills concerning deepfakes include:

- The Protecting Consumers from Deceptive AI Act (2024) would order NIST to develop standards for identifying and labeling AI-generated content and require developers to include machine-readable disclosures within content. This bill was co-sponsored by a bipartisan group in the House only.[1794]
- The Preventing Deepfakes of Intimate Images Act (2024) goes further, seeking to criminalize the disclosure of sexually explicit deepfake content. House bill only.[1795]
- The No Artificial Intelligence Fake Replicas And Unauthorized Duplications Act (No AI FRAUD Act)[1796] seeks to prevent the unauthorized use of any individual's "likeness and voice." House bill only.

[1791] *Id.*; AI Foundation Model Transparency Act of 2023, H.R.6881, 118th Cong. (2023).

[1792] Bommasani et al., *Foundation Model Transparency Reports*, arXiv, (Feb. 26, 2024), https://doi.org/10.48550/arXiv.2402.16268.

[1793] Press Release, Brian Schatz, U.S. Senate, Schatz, Kennedy Introduce Bipartisan Legislation To Provide More Transparency On AI-Generated Content, (Oct. 24, 2023), https://www.schatz.senate.gov/news/press-releases/schatz-kennedy-introduce-bipartisan-legislation-to-provide-more-transparency-on-ai-generated-content; Press Release, Tom Kean, Jr., House of Representatives, Kean Introduces Bill to Provide More Transparency on AI-Generated Content (Nov. 27, 2023), https://kean.house.gov/media/press-releases/kean-introduces-bill-provide-more-transparency-ai-generated-content.

[1794] Press Release, Anna G. Echoo, House of Representatives, *Rep. Eshoo Introduces Bipartisan Bill to Label Deepfakes* (Mar. 21, 2024), https://eshoo.house.gov/media/press-releases/rep-eshoo-introduces-bipartisan-bill-label-deepfakes.

[1795] Press Release, Joe Morelle, House of Representatives, Congressman Joe Morelle Takes Action to End AI Generated Deepfake Pornography (Jan. 16, 2024), https://morelle.house.gov/media/press-releases/congressman-joe-morelle-takes-action-end-ai-generated-deepfake-pornography.

[1796] Press Release, Maria Elvira Salazar, House of Representatives, Salazar Introduces the No AI Fraud Act (Jan. 10, 2024), https://salazar.house.gov/media/press-releases/salazar-introduces-no-ai-fraud-act.

- The Nurture Originals, Foster Art, and Keep Entertainment Safe Act (NO FAKES Act)[1797] would also protect against the unauthorized use of someone's "voice and visual likeness" using generative AI. A bipartisan group of senators introduced the legislation.
- Senator Richard Durbin and Representative Alexandria Ocasio-Cortez introduced identical bills in the Senate and House entitled the Disrupt Explicit Forged Images and Non-Consensual Edits Act (DEFIANCE Act) of 2024. This legislation has perhaps the highest likelihood of any deepfake bill of being adopted because of its prominent bipartisan and bicameral sponsors. It would create a federal civil right of action for victims of nonconsensual, sexually explicit deepfakes.[1798]

Whether any of these bills ultimately passes, addressing deepfakes will doubtless remain a recognized priority in any future AI legislation.

e) Artificial Intelligence Environmental Impacts Act of 2024

In February 2024, Senator Ed Markey and Representative Anna Eshoo introduced the Artificial Intelligence Environmental Impacts Act to establish national standards for measuring AI's environmental impact. The legislation would direct NIST to develop standards to measure and report the full range of AI's environmental impacts. It would also create a voluntary framework for AI developers to report the impact their models have on the environment. It would mandate an interagency study to investigate and measure both the positive and negative environmental impacts of AI. This bill has the backing of a host of corporate and civil society groups, counting endorsements from Hugging Face, Public Citizen, Sierra Club, Greenpeace USA, and the Center for AI and Digital Policy, among others.[1799]

f) Generative AI Copyright Disclosure Act of 2024

Representative Adam Schiff introduced the Generative AI Copyright Disclosure Act in April 2024. The bill, introduced in the House only, seeks to require AI model developers to file with the Register of Copyrights a summary of all copyrighted works used in their training datasets. This brief bill would not alter existing copyright law nor frustrate the use of copyrighted materials to train models.[1800] Nevertheless,

---

[1797] No AI FRAUD Act, H.R.6943, 118th Cong. (2024); https://www.congress.gov/bill/118th-congress/house-bill/6943/text?s=1&r=1&q=%7B%22search%22%3A%22no+fakes+act%22%7D; Chris Coons, U.S. Senate, One-page summary of NO FAKES Act (Oct. 12, 2023), https://www.coons.senate.gov/imo/media/doc/no_fakes_act_one_pager.pdf.

[1798] Richard J. Durbin, U.S. Senate, One-pager summarizing The DEFIANCE Act of 2024 (Jan. 30, 2024), https://www.durbin.senate.gov/imo/media/doc/defiance_act_of_2024.pdf; Press Release, Alexandra Ocasio-Cortez, House of Representatives, Rep. Ocasio-Cortez Leads Bipartisan, Bicameral Introduction of DEFIANCE Act to Combat Use of Non-Consensual, Sexually-Explicit "Deepfake" Media (Mar. 7, 2024), https://ocasio-cortez.house.gov/media/press-releases/rep-ocasio-cortez-leads-bipartisan-bicameral-introduction-defiance-act-combat.

[1799] Press Release, Ed Markey, U.S. Senate, Markey, Heinrich, Eschoo, Beyer Introduce Legislation to Investigate, Measure Environmental Impacts of Artifical Intelligence (Feb. 1, 2024), https://www.markey.senate.gov/news/press-releases/markey-heinrich-eshoo-beyer-introduce-legislation-to-investigate-measure-environmental-impacts-of-artificial-intelligence.

[1800] Press Release, Adam Schiff, House of Representatives, Rep. Schiff Introduces Groundbreaking Bill to Create AI Transparency Between Creators and Companies (Apr. 9, 2024),

the step toward greater transparency has garnered support from entertainment industry organizations and unions, including the Recording Industry Association of America, Professional Photographers of America, Directors Guild of America, and the Screen Actors Guild-American Federation of Television and Radio Artists.[1801]

g) Future of Artificial Intelligence Innovation Act of 2024

Also in April 2024, a bipartisan group of senators on the Commerce Committee introduced the Future of Artificial Intelligence Innovation Act. The legislation would formally bless NIST's extant AI Safety Institute and Safety Institute Consortium (AISIC) and authorize it to continue developing voluntary guidelines and standards in tandem with the private sector to promote long-term innovation in AI in the United States. The bill seeks to create testbed programs, led by NIST and other science agencies in public-private partnership with industry, to accelerate innovation, particularly in materials science and advanced manufacturing. The proposed law would direct these agencies to make curated datasets available for public use so as to accelerate private sector advances in AI applications. The bill would also encourage US government agencies like NIST to forge international alliances on AI standards, research, and development.[1802]

h) ENFORCE Act

Finally, in May 2024, a bipartisan coalition of national security-minded members introduced the Enhancing National Frameworks for Overseas Restriction of Critical Exports Act (ENFORCE Act).[1803] This bill tackles the potential national security implications of AI by granting the Commerce Department the authority to impose export controls or an export licensing regime for AI systems or models. The bill would extend the Commerce Department's authority beyond its existing limits to cover "any software and hardware implementation of artificial intelligence," including model weights. Currently, Commerce can restrict the export of only semiconductors used to create AI systems – not the systems themselves.[1804] The introducing members' press release expressly calls out exports to China as jeopardizing national security,

---

https://schiff.house.gov/news/press-releases/rep-schiff-introduces-groundbreaking-bill-to-create-ai-transparency-between-creators-and-companies.

[1801] Nick Robins-Early, *New bill would force AI companies to reveal use of copyrighted art*, THE GUARDIAN (Apr. 9, 2024), https://www.theguardian.com/technology/2024/apr/09/artificial-intelligence-bill-copyright-Article.

[1802] *Cantwell, Young, Hickenlooper, Blackburn Introduce Bill to Ensure U.S. Leads Global AI Innovation*, U.S. SENATE COMMITTE ON COMMERCE, SCIENCE & TRANSPORTATION (Apr. 18, 2024), https://www.commerce.senate.gov/2024/4/cantwell-young-blackburn-hickenlooper-introduce-bill-to-ensure-u-s-leads-global-ai-innovation; Future of Artificial Intelligence Innovation Act of 2024, H.R., 118th Cong. (2024). https://www.commerce.senate.gov/services/files/E60CB738-9C67-4FD2-8F28-26B5D5EC33BE.

[1803] ENFORCE Act, H.R.8315, 118th Cong. (2024), https://www.congress.gov/bill/118th-congress/house-bill/8315/text.

[1804]Press Release, Foreign Affairs Comm., Bipartisan Coalition Introduces Monumental Bill Giving Admin Authority to Export Control Advanced AI Systems (May 10, 2024), https://foreignaffairs.house.gov/press-release/bipartisan-coalition-introduces-monumental-bill-giving-admin-authority-to-export-control-advanced-ai-systems/#:~:text=The%20ENFORCE%20Act%20would%20allow,do%20not%20threaten%20national%20security.

and notes that without the bill, not only can top US companies sell systems that threaten national security, but American researchers can work in Chinese AI labs.[1805]

3) Conclusion

Most of the pending legislative proposals in the Congress cover a different specific issue or were crafted in immediate response to recent AI-related news and, as a result, may not have enough momentum to become law in 2024. Even so, there are certain commonalities that emerge across these bills that may endure in future Congressional sessions. The protection of sensitive data categories, the use of AI in critical decision-making, and its impact on vulnerable populations are some of the enduring concerns that have emerged across proposed US legislation thus far. These bills acknowledge, and perhaps will reify, the conceptual distinction between deployers and developers. There is also a clear solicitude for the harms perpetrated by deepfakes and other unauthorized use of content. Finally, legislators seem equally concerned about preserving the benefits of AI and American leadership in AI as they are about its risks. Most statements accompanying proposed legislation reveal a reluctance to overregulate or stifle innovation. Put another way, there does not appear to be much appetite in the US to ameliorate AI risks with a new federal agency or sweeping new rules. Instead, the more likely legislative synthesis appears to be through disclosure, transparency, and technical guidelines. If more government resources are needed, the preference seems to be to bolster existing agencies with new legal authority and funding.

Perhaps the greatest takeaway from the pending legislation has been lawmakers' already substantial and growing engagement with interested civil society participants—including many nontraditional players in technology regulation—in forging alliances and getting endorsements for these proposed laws. This suggests coalition-building in this emerging area is already in full swing and may prove useful in hastening and ultimately pushing final legislation through.

5.3.3. State regulatory initiatives in the US

A vacuum in US federal regulation can prompt state governments to fill the void with their own laws. In the case of AI, the comprehensive data privacy laws enacted by many states contain various provisions that could have an impact on different aspects of the training and deployment of generative AI models. But many states started to address generative AI more directly through the formation of task forces, the ordering of reports and future guidance, and the introduction of a variety of bills.

5.3.3.A. State legislation and other initiatives

The National Conference of State Legislatures (NCSL) published a report in August 2023 aimed at helping state legislators get up to speed and understand AI best practices, key risks, and common definitions in

[1805] *Id.*; Michael McCaul, *Enhancing National Frameworks for Overseas Critical Exports Act (ENFORCE Act)*, https://foreignaffairs.house.gov/wp-content/uploads/2024/05/ENFORCE-Act-Bill-Summary.pdf, (last visited Jun. 29, 2024).

order to draft future laws.[1806] Many states have drafted new laws, principally on deepfakes and particularly those that feature nonconsensual sexual images.

1) State laws addressing deepfakes

According to the National Conference of State Legislatures (NCSL), nearly every state legislature is considering AI-related bills, with most of those bills addressing concerns about deepfakes.[1807] Though there are hundreds of bills proposed or pending, there are essentially two categories of deepfake laws: one targeting non-consensual, sexually-explicit deepfakes, and a second targeting the use of deepfakes in elections. As of early July 2024, Multistate.ai counts 29 states have enacted laws addressing the former and 18 states addressing the latter.[1808] Among the states that have criminalized the non-consensual dissemination or publication of sexually-explicit deepfakes are Virginia,[1809] New York,[1810] and Texas.[1811] Other states like California[1812] and Illinois[1813] grant victims a private right of action. A few states like Minnesota offer both criminal penalties and civil recourse for victims.[1814] Many states such as Florida,[1815] South Dakota,[1816] and Louisiana,[1817] have not banned sexually explicit deepfakes of adults, but have updated their child pornography / CSAM laws to proscribe possessing or creating deepfakes of minors.[1818] Although most CSAM laws require the minor to be an identifiable person, and continue to require this for deepfakes, some new state laws like Florida's and South Dakota's prohibit even virtual CSAM deepfakes that are not necessarily linked to any identifiable person.

The rising use of deepfake voice and video imitations in the early goings of the 2024 election campaign has spurred the introduction of legislation targeting political figure deepfakes. These bills either ban, or mandate

---

[1806] NCSL, *Approaches to Regulating Artificial Intelligence: A Primer* (Aug. 10, 2023), https://www.ncsl.org/technology-and-communication/approaches-to-regulating-artificial-intelligence-a-primer.

[1807] NCSL's tracker counts "at least" 40 states have proposed laws in 2024 alone. NCSL, Artificial Intelligence 2024 Legislation (last updated Jun. 3, 2024), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation.

[1808] *Deepfakes & Synthetic Media*, MULTISTATE.AI, https://www.multistate.ai/deepfakes-synthetic-media (last updated July 11, 2024).

[1809] Va. Penal § 18.2-386.2 (2019), https://lis.virginia.gov/cgi-bin/legp604.exe?191+sum+HB2678.

[1810] N.Y. SB 1042A, amending N.Y. PENAL LAW § 245.15 (2024), https://www.nysenate.gov/legislation/bills/2023/S1042/amendment/A.

[1811] Tex. SB 1361, Unlawful Production or Distribution of Certain Sexually Explicit Videos, Tex. Penal tit. 5 § 21.165 (2024), https://capitol.texas.gov/tlodocs/88R/billtext/html/HB02700H.htm.

[1812] Cal. AB-602 (2019), Depiction of individual using digital or electronic technology: sexually explicit material, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602.

[1813] Ill. Pub. Act 103-0571 (2023), https://www.ilga.gov/legislation/publicacts/fulltext.asp?Name=103-0571.

[1814] Minn. HF 1370 (2023), https://www.revisor.mn.gov/bills/text.php?session=ls93&number=HF1370&session_number=0&session_year=2023&version=list&format=pdf.

[1815] Fla. SB-1680 (2024), https://flsenate.gov/Session/Bill/2024/1680/BillText/er/HTML.

[1816] S.D. SB 79 (2024), https://legiscan.com/SD/text/SB79/id/2916028.

[1817] La. SB-175 (2023), https://legis.la.gov/Legis/ViewDocument.aspx?d=1333325.

[1818] Madyson Fitzgerald, *States race to restrict deepfake porn as it becomes easier to create*, STATELINE (Apr. 10, 2024), https://stateline.org/2024/04/10/states-race-to-restrict-deepfake-porn-as-it-becomes-easier-to-create/.

affirmative disclosure, whenever AI is used to create material to influence an election.[1819] Some states like California[1820] and Texas[1821] already had legacy laws on the books banning deceptive or doctored media involving candidates for elected office. Others like Indiana,[1822] Colorado,[1823] and Wisconsin[1824] adopted new laws in 2024 specifically responding to AI-generated political deepfakes. Colorado's and New Hampshire's[1825] laws go further by also granting a private right of action.

In March 2024, Tennessee became the first state to prohibit the unauthorized use of deepfakes of any person's voice, with passage of the Ensuring Likeness, Voice, and Image Security (ELVIS) Act of 2024.[1826] The ELVIS Act amended the 1984 Personal Rights Protection Act which created a property right in the use of a person's name, photograph, or likeness. The amendment added a person's voice (defined to include actual or simulations of that voice). The ELVIS Act allows aggrieved individuals to protect these property rights by suing those who unlawfully publish, distribute, or transmit an individual's voice, name, photograph, or likeness without the person's authorization. The private right of action further extends to anyone who distributes, transmits, or otherwise makes available an algorithm, software, tool, or other technology or service whose "primary purpose or function" is to enable the unauthorized distribution –so long as that person had knowledge that the content was not authorized.[1827] The precise contours of this primary purpose or function test are unclear.[1828] And though the law contains an exception for federal fair use and for protected First Amendment expression, it will likely fall to the courts to draw those lines.

2) State laws targeting AI more broadly: The example of Colorado

States have not stopped with bills on deepfakes and appear imminently poised to consider legislation on AI topics beyond deepfakes. The first mover in this respect is Colorado. On May 17, 2024, Colorado passed the first state law targeting algorithmic discrimination: the Consumer Protections for AI Act (SB 24-205).[1829] Colorado SB 205 borrows from the EU's AI Act by training its sights on "high-risk" AI systems only, albeit a narrower set of risks: It covers any system used that has a material or substantial effect on decisions relating to education, employment, credit/lending, essential governmental services, healthcare, housing, insurance, or legal services. The law grants consumers a right to an explanation of any adverse

---

[1819] *See* Tracker: State Legislation on Deepfakes in Elections, PUBLIC CITIZEN, https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/.

[1820] Cal. AB-972 (2023), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB972.

[1821] Tex. S.B. 171 (2019), https://legiscan.com/TX/text/SB751/2019.

[1822] Ind. HB-1133 (2024), https://iga.in.gov/pdf-documents/123/2024/house/bills/HB1133/HB1133.06.ENRS.pdf.

[1823] Candidate Election Deepfake Disclosures, Col. HB24-1147 (2024), https://leg.colorado.gov/bills/hb24-1147.

[1824] 2023 Wis. Act 123 (2024), https://docs.legis.wisconsin.gov/2023/related/acts/123.

[1825] N.H. HB1432-FN (2024), https://www.gencourt.state.nh.us/bill_Status/billinfo.aspx?id=1239.

[1826] Tenn. HB-2091 (2024), https://publications.tnsosfiles.com/acts/113/pub/pc0588.pdf; *see also* Bill Summary, Tenn. Gen. Assembly Legislation, https://wapp.capitol.tn.gov/apps/BillInfo/Default.aspx?BillNumber=HB2091.

[1827] *Id.*

[1828] Jesse Feitel, Nicolas A. Jampol & James Rosenfeld, *Tennessee, All Shook Up Over AI-Generated Voice Replicas, Passes ELVIS Act* (Apr. 8, 2024), https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2024/04/tennessee-elvis-act-ai-voice-replica.

[1829] Col. SB 24-205 (2024), https://leg.colorado.gov/sites/default/files/2024a_205_signed.pdf.

consequential decision, the right to correct inaccurate information used by the system, and to appeal the model's decision for human review, if feasible.

The law imposes a duty on *both* developers and deployers to take reasonable care against discrimination in those areas. It does so by adopting certain mitigations, including public disclosures about known or reasonably foreseeable risks, documentation describing the measures taken to counteract those risks, a risk management policy and program, impact assessments, and an affirmative notice requirement to the state attorney general if the system has caused or is reasonably likely to cause discrimination. The attorney general has rulemaking authority and exclusive jurisdiction for enforcing this law, meaning there is no private right of action. In an enforcement action, developers or deployers may plead an affirmative defense if (a) they discover and cure the violation in accordance with public feedback, red teaming, or an internal review process and (b) they have implemented and maintained a program that is in compliance with NIST's *AI RMF* or another nationally or internationally recognized risk management framework for AI.[1830]

Though Colorado's law does not take effect until 2026, it is the most substantive state law regulating AI yet, and it may cause a domino effect with other states. It may also not. A very similar legislative effort in Connecticut (Conn. S.B. 2) derailed when it passed one house, but the Connecticut governor unexpectedly threatened to veto it, on grounds that it was premature and better suited for federal action.[1831] Colorado's governor signed the law notwithstanding similar reservations.[1832]

FIGURE 48. Examples of adopted state regulations covering AI

State	Action	Summary
California	Risk Assessment and Automated Decision-making Technology Regulations (2023) EO N-12-23 (2023)	Consumer Privacy Protection Agency (CPPA) regulations would require businesses that use automated decision-making technology to notify consumers about how the business intends to use it, so that consumers can decide whether to opt-out or access more information. Businesses would also need to complete risk assessments. Governor's executive order commissioned a draft task force report (2023) and risk assessment for critical energy infrastructure, to be completed in 2024
Colorado	SB24-205, Consumer Protections for Artificial Intelligence (2024)	Requires developers and deployers of high-risk AI systems in consequential decision-making to use reasonable care to avoid algorithmic discrimination, including measures such as notice, documentation, disclosures, and impact assessments.

[1830] Col. S.B. 24-205 (2024), https://leg.colorado.gov/sites/default/files/documents/2024A/bills/2024a_205_enr.pdf.
[1831] Mallory Culhane, *Two unlikely states are leading the charge on regulating AI*, POLITICO (May 15, 2024), https://www.politico.com/news/2024/05/15/ai-tech-regulations-lobbying-00157676.
[1832] Jared Polis, *Letter to the Honorable Members of the Colorado General Assembly* (May 17, 2024), https://aboutbgov.com/bd7s.

Connecticut	S.B. No. 1103, An Act Concerning AI, Automated Decision-Making and Personal Data Privacy (2023) Privacy Act (2023)	Established a state Office of AI and a task force to study AI and develop an AI bill of rights Provides consumers the right to opt out of profiling in furtherance of solely automated decisions.
Florida	CS/HB 919 - AI Use in Political Advertising (2024)	Requires political ads and electioneering communications to disclose use of AI through specified disclaimer
Illinois	AI Video Interview Act (2020, amended 2022)	Employers must provide notice of their use of AI in job interviews and obtain applicants' consent.
New York	Automated Employment Decisions Tools Act, L.L. 2021/144 (2021)	New York City employers must notify job candidates about the use of AI tools in hiring, and they must conduct bias audits of AI-enabled tools used for employment decisions.
Utah	AI Policy Act (2024)	Expands existing consumer protection law on unfair and deceptive practices to suppliers of generative AI. Creates Office of AI Policy and a regulatory AI analysis program. Requires disclosure when an individual interacts with AI in a regulated occupation.
Texas	AI Advisory Council (2023)	Study and monitor AI developed or employed by Texas state agencies.
Tennessee	ELVIS Act (2024)	Prohibit the unauthorized use of deepfakes, including voice replicas. Offers private right of action to aggrieved individuals to enforce their rights against infringement, including right to pursue any person who distributes, transmits, or otherwise makes available an AI system if the the primary purpose is the production of unauthorized content.
Vermont	AI Task Force (2020)	Authored report and established Division of AI, which conducts a yearly inventory of the use of AI within government
Wisconsin	2023 Act 123 (2024)	Any audio political communication using generative AI shall include, both at the beginning and at the end of the communication, the words "Contains content generated by AI." Any video political communication using generative AI shall include "This content generated by AI."

3) Pending state laws: The example of California

Many states have proposed laws on AI with great alacrity. These bills face varying prospects of ultimate passage but, if nothing else, they will get speedier resolution in state legislatures relative to their federal

counterpArticle And because many states have the same political party ruling both executive and legislative branches, there is a higher probability of laws being passed. The paradigmatic example of this is California. As of May 2024, many AI-related bills are ripening in California and will be up for a vote as early as this August.[1833] Since California is the home of many leading AI developers and has long been at the vanguard of technology regulation among the US states, it is worth examining California's proposed laws more closely. Although California is not the first mover on AI, it may become the most important and the industry's de facto US regulator, particularly if some of its most ambitious proposals (out of some 30 advanced to date) pass.[1834]

On the mild end of the spectrum is California's Senate Bill 896, the Generative AI Accountability Act.[1835] This bill would require state agencies to produce a report examining beneficial uses of AI and also to notify consumers when they interact with AI systems utilized by the state agencies. In the middle of the spectrum is Assembly Bill 2930, Automated Decision Tools, which would mandate impact assessments by companies which use automated decision-making tools, as well as notice to subjects of consequential decisions taken by such tools.[1836] This bill is redolent of Colorado SB 205 and shares the same intent – to ward off or keep algorithmic discrimination at bay.

Most sweeping is SB 1047**,** the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.[1837] California SB 1047 has passed one house of California's legislature and, if duly enacted, would impose a battery of compliance measures on developers of AI models that present "hazardous capabilities." Those measures include:

- administrative, technical, and physical cybersecurity protections to prevent unauthorized access, misuse, or unsafe modification of the covered model;
- a full shutdown "kill switch" and the capability to promptly enact a full shutdown of the covered model and copies and derivative models;
- the incorporation of all available NIST safety guidance and other voluntary best practices into law; and
- the implementation of a written and separate safety and security protocol that provides reasonable assurance that the model has sufficient safeguards to prevent an unreasonable risk of critical harms.

The bill defines "covered models" as the federal government defined "frontier models"—models whose training requires computational resources in excess of 10^{26} FLOPS—but also appends a $100 million cost

---

[1833] Cecilia Kang, *States Take Up A.I. Regulation Amid Federal Standstill,* N.Y. TIMES (June 10, 2014), https://www.nytimes.com/2024/06/10/technology/california-ai-regulation.html.

[1834] Jennifer Huddleston, *AI Could Become the Next Victim of the 'Sacramento Effect'*, REASON (June 7, 2024), https://reason.com/2024/06/07/ai-could-become-the-next-victim-of-the-sacramento-effect/ (decrying California's regulatory stringency as a "Sacramento effect" similar to the "Brussels effect").

[1835] Generative Artificial Intelligence Accountability Act, SB-896 (Cal. 2024), https://legiscan.com/CA/text/SB896/2023.

[1836] Automated decision tools, AB-2930 (Cal. 2024), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2930.

[1837] Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, SB-1047 (Cal. 2024), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047; *see also* Joshua Turner and Nicol Turner Lee, *Can California fill the federal void on frontier AI regulation?*, BROOKINGS (June 4, 2024), https://www.brookings.edu/articles/can-california-fill-the-federal-void-on-frontier-ai-regulation/.

requirement for that training (based on average current prices of cloud compute). The bill specifies in some detail what constitutes a "hazardous capability" and how a developer might find "reasonable assurance," through testing, that a model does not present any such hazards.. A "hazardous capability" means the capability to create a CBRN weapon, cause $500 million in cyberattack harm, or autonomously cause $500 million in damages, if that damage also causes bodily harm or harm to property. Reasonable assurance testing must ensure that covered models do not have or "come close to" having hazardous capabilities. As such, the bill leaves many key terms vague and would harden into law voluntary NIST guidance and inchoate industry practices. Moreover, compliance with these prophylactic measures must be undertaken *before* frontier models are even trained. And though none of these requirements applies to developers of "derivative models"—which might spare smaller deployers of AI models—developers must ensure that they do not enable production of such derivative models with hazardous capabilities.

Finally, though this list of compliance requirements would not disturb *noncovered* models that do not meet the computing power or dollar cost thresholds, it would require weaker models to implement the applicable NIST safety guidance and other industry best practices.

To administer these multifarious requirements, California would create a new Frontier Model Division within the existing Department of Technology and charge that division with reviewing annual certifications by developers of non-frontier models, the safety and security protocols of frontier models, safety incident reports, and issuing guidance, standards and best practices to prevent unreasonable risks from frontier models.

As of this writing, the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act is still undergoing revisions in both chambers in California, and so its final form (if any) will doubtless differ from the above. Nevertheless, the bill has already begun to garner criticism for many reasons. First, it does not offer any exemption for open-source models (aside from the full shutdown capability), which would make open-source developers culpable for the wrongful actions of others outside of their control – severely penalizing open-source development.[1838] More significantly, by targeting "hazardous capabilities" of the models themselves, SB 1047 aims to deter AI model builders from creating technology that could potentially cause future harm, regardless of the model's primary design or the intentions at the time of training. This approach shifts responsibility away from downstream applications that may actually pose harm or a hazard. Simply put, the bill could be seen as a Procrustean attempt to regulate the technology as a whole, rather than regulating applications, a strategy which could stunt or neuter all of the beneficial uses of AI solely because of what harm a model might ever do.[1839] On the pecuniary side, the bill would impose steep civil penalties on a nascent industry which still lacks clear federal guidance and where even industry norms and best practices are only beginning to coalesce. SB 1047 is not without its supporters, though. Some proponents celebrate the bill's numerous safeguards and argue this bill does not go far enough to

[1838] *A statement in opposition to California SB 1047*, AI ALLIANCE, https://thealliance.ai/core-projects/sb1047 (last visited July 15, 2024).

[1839] *See* Andrew Ng, *Issue 252*, THE BATCH (June 5, 2024), https://www.deeplearning.ai/the-batch/issue-252/. It is speculated that this bill is in part motivated by prominent AI doomsayers, as well as those with ties to the effective altruism movement.

forestall the present public safety risks or prophecies of AI's future existential risk.[1840] The sponsor of the bill has rejoined that if Congress will not act, the states must.[1841]

TABLE 49 Examples of pending legislation covering AI.

State	Action	Summary
California	SB 1047, Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (proposed 2024)	Would create a new office in the California Department of Technology called the Frontier Model Division, tasked with strengthening AI enforcement and reviewing annual certifications by powerful frontier model developers. Would require developers to conduct pre-deployment safety testing, make positive safety determinations, and implement cybersecurity protections, including the capability to fully shut down the model.
	AB 2930, Automated Decision Tools (proposed 2024)	Would require deployers (e.g., employers) to perform an impact assessment before using automated decision tools and to notify affected subjects in advance that such systems would be used. It would also prohibit deployers from using such tools to make consequential decisions that result in algorithmic discrimination.
	AI Accountability Act (proposed 2024)	Would require California agencies to produce a report examining beneficial uses of AI by the state. Would also require state agencies to notify users when interacting with AI.
Colorado	HB24-1147, Candidate Election Deepfake Disclosures (proposed 2024)	Would impose civil penalties for distributing deepfakes regarding a candidate for elective office.
Connecticut	S.B. 2, An Act Concerning Artificial Intelligence (proposed 2024)	Would require developers and deployers of high-risk AI systems in consequential decision-making to use reasonable care to protect consumers from algorithmic discrimination. Would also penalize the creation and dissemination of nonconsensual deepfake content and require AI-generated content to contain digital watermarks.

[1840] Gabriel Weil, *The Pros and Cons of California's Proposed SB-1047 AI Safety Law*, LAWFARE (May 8, 2024), https://www.lawfaremedia.org/article/california-s-proposed-sb-1047-would-be-a-major-step-forward-for-ai-safety-but-there-s-still-room-for-improvement.

[1841] Shirin Ghaffary, *Silicon Valley Is On Alert Over a Proposed AI Bill in California*, BLOOMBERG (June 6, 2024), https://www.bloomberg.com/news/newsletters/2024-06-06/silicon-valley-is-on-alert-over-an-ai-bill-in-california.

Mass.	H1873, An Act Preventing a Dystopian Work Environment (proposed 2023) SB 31, An Act drafted with the help of ChatGPT to regulate generative artificial intelligence models like ChatGPT (proposed 2023)	Would require employers to provide workers notice about the use of automated decision systems/algorithms and the right to request information about such systems. Would require companies operating large-scale generative AI models to adhere to certain operating standards, such as: not engaging in discrimination or bias; preventing plagiarism through watermarking of AI-generated content; implementing reasonable security measures to protect personal data used to train the model; and obtaining consent from individuals before collecting, using, or disclosing their data. Companies would also be required to perform regular risk assessments and register with the state attorney general.
New York	Digital Fairness Act (proposed 2023)	Would require AI impact assessments, prevent discriminatory practices, and regulate the use of biometric data.
Texas	HB4695 (proposed 2023)	Would prohibit the use of AI for counseling and therapy without approval by Texas commission and supervision by licensed professionals.
Vermont	H114 (proposed 2023)	Would restrict use of automated decision systems for employment-related decisions.

5.3.3.B. The interplay between state and federal initiatives

As state regulatory initiatives march ahead of federal ones, the US may end up with a legal environment that varies by state. Or, if a federal law is enacted, it could either create overlapping jurisdictions between state and federal law or land on a single national uniform standard. This all depends on how state and federal laws interact. There are a range of possible outcomes with layers of nuance, and these scenarios may unfold in ways that are hard to predict.

That said, the history of the data privacy laws could be instructive. As more states pass laws in a given field, pressure rises on the federal government to do the same. Much of this pressure comes from civil society groups which see such laws as offering desirable protections that should be extended nationwide. Just as importantly, the proliferation of state laws can also lead to calls for federal action from industry proponents who see a growing patchwork of state laws as increasing operational complexity and compliance costs.

A patchwork regime may not necessarily be problematic; in the eyes of some, it fulfills the Brandeis-ian vision of states as "laboratories of democracy." It could also be a fertile testing ground, providing the federal government with valuable information on what works and what does not. However, the entrenchment of state laws regulating a certain field can also lead to significant complexity and conflict when Congress sets out to adopt a federal law in that same field.

There are two broad possibilities for how federal-state tension can play out. One option, preemption, would have federal law become the exclusive law regulating the field, displacing state laws that also seek to regulate all or part of the field. The other option is for federal law to merely provide a "floor," setting minimum standards in the field but leaving states free to enact more stringent laws. The former, of course, is generally favored by industry since it simplifies the compliance environment, while the latter is generally preferred by civil society groups and state constituents, who want to leave room for states to impose tougher standards.

Disagreements over preemption can be fatal to otherwise promising legislation. In one particularly relevant example, the most serious attempt so far at enacting a federal data privacy law, the introduction of the ADPPA in the 117th Congress, foundered on the issue of preemption. Congressional representatives from states like California, that already have strong data privacy laws and enforcement agencies, balked at proposals for the ADPPA to preempt state privacy laws.[1842] The preemption hurdle appears to have been cleared with the freshly proposed American Privacy Rights Act of April 2024, which augurs well for any future federal AI legislation.[1843]

KEY TAKEAWAYS

- The United States lacks a comprehensive legal framework to govern artificial intelligence at the federal and the state levels. Existing legal frameworks are fragmented and encompass state data protection laws, federal intellectual property laws, and principles of general liability.

- The training of AI models using scraped, publicly available content on the internet is generally exempt from extant state data protection laws. This is because web-scraped data are typically regarded as "publicly available" information, meaning it is information that a business has a reasonable basis to believe has already been made lawfully accessible to the public.

- In the realm of copyrights, significant lawsuits have been filed over the past 18 months by authors, artists, and media companies. These lawsuits allege that AI companies have trained generative AI models on copyrighted works without the companies obtaining permission from or providing compensation to the creators of the copyrighted material. None of these pending copyright cases

[1842] *See* John D. McKinnon, *Data-Privacy Bill Advances in Congress, but States Throw Up Objections*, WALL ST. J. (Jul. 20, 2022), https://www.wsj.com/articles/data-privacy-bill-advances-in-congress-but-states-throw-up-objections-11658347139.

[1843] *Committee Chairs Rodgers, Cantwell Unveil Historic Draft Comprehensive Data Privacy Legislation*, Committee on Energy and Commerce (Apr. 7, 2024), https://energycommerce.house.gov/posts/committee-chairs-rodgers-cantwell-unveil-historic-draft-comprehensive-data-privacy-legislation; Jedidiah Bracy, *New draft bipartisan US federal privacy bill unveiled,* IAPP (Apr. 7, 2024) https://iapp.org/news/a/new-draft-bipartisan-us-federal-privacy-bill-unveiled/.

has yet been fully adjudicated on their merits, and the outcome of these claims will likely depend on whether the use of the copyrighted material can be deemed "fair use." The issue of whether outputs generated by generative AI tools can infringe on copyrights also remains unresolved. Some generative AI providers now offer indemnification to potentially affected users, indicating the companies consider the risk of future litigation to be low. Uncertainties also exist regarding the eligibility of AI-generated content for copyright or patent protection.

- The debate is intense regarding whether companies that develop generative AI models and systems should be held liable under U.S. law for harms caused by their tools. Companies like OpenAI and Google include disclaimers with their chatbot systems to alert users about the potential for inaccurate or misleading outputs, although the legal effectiveness of these disclaimers remains uncertain. Questions are also still unresolved about whether AI-generated speech is protected under the First Amendment and exempt from liability under Section 230 of the Communications Decency Act. Furthermore, it is not yet established that product liability laws apply to this sphere, despite some scholars advocating for this position.

- In response to the rapid release of increasingly sophisticated generative AI systems, U.S. federal authorities have initially prioritized dialogue with major AI developers and the development of non-binding rules and standards. In October 2022, the White House Office of Science and Technology Policy unveiled a "Blueprint for an AI Bill of Rights." This document outlined five principles and associated practices to guide the private sector in the design, use, and deployment of AI, aiming to protect public rights. Furthermore, the White House engaged the AI industry through informal dialogue and collaboration, securing voluntary commitments from leading AI companies in July and September 2023.

- Simultaneously, the Biden administration has urged independent regulatory agencies to leverage their existing legal authority to monitor the AI ecosystem. The Federal Trade Commission (FTC), in particular, has shown its intent to be a robust regulator of the industry, even without new, specific legal authority to regulate AI. The FTC contends that its current powers can be used to protect consumers and competition from harms caused by generative AI companies. In April 2023, the Consumer Financial Protection Bureau (CFPB), along with other federal agencies, issued a joint statement pledging to use their existing authority to combat AI-based bias and discrimination in their respective regulatory domains. The Equal Employment Opportunity Commission (EEOC) has also issued technical guidance on how AI's use in hiring and evaluating employees could be affected by anti-discrimination law requirements.

- After initially focusing on the development of non-binding standards, the Biden administration has transitioned to adopting an ambitious Executive Order. The October 2023 Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence represents the most significant legal and policy action on AI in the U.S. to date. This executive order announces eight guiding principles and policy priorities for the administration and instructs various executive

branch departments to issue reports, studies, guidelines, plans, and take other appropriate actions under existing legal authority. It also establishes mechanisms and bodies through which different entities within the federal government can coordinate on AI issues and solicit input from relevant stakeholders outside the government. These efforts could form the factual or policy basis for future legally binding actions, such as those by Congress or regulatory agencies. Notably, the executive order includes precise definitions of technical terms, introducing the category of "dual-use foundation model," and provides pragmatic directives to federal agencies. This approach differs significantly from the EU AI Act, which is primarily a legislative document that focuses on legal concepts and delegates the technical details to the authorities responsible for implementation.

- Meanwhile, the National Institute of Standards and Technology (NIST), which is a non-regulatory agency, has published several non-binding documents based on its AI Risk Management Framework, addressing the risks and cybersecurity best practices associated with generative AI and dual-use foundation models. Although there is no formal enforcement mechanism for these standards, AI companies are expected to follow them with a high degree of commitment. This expectation is especially strong because these standards were developed through joint collaboration among NIST, industry representatives, and various stakeholders. To further enhance government-industry cooperation on AI safety, the Biden administration has established the AI Safety Institute (AISI) under NIST's auspices to develop guidelines and standards for AI measurement and policy.

- Overall, the current strategy of the Biden administration appears to lean towards a form of "encouraged self-regulation." By securing voluntary commitments that it cannot verify compliance with and having NIST develop non-mandatory standards in collaboration with the industry, the federal administration mostly relies on the goodwill of the private sector and their desire to maintain their reputations by effectively adhering to certain agreed-upon rules.

- In the absence of a federal legal framework, several legislative proposals have been submitted in Congress. Some are broad frameworks designed to serve as a policy foundation for future discussions and legislation, while others are targeted bills addressing specific, known dangers of AI. Among the boldest initiatives is the Blumenthal-Hawley Bipartisan Framework for U.S. AI Act, which proposes the establishment of a licensing regime administered by an independent oversight body. However, neither this framework nor any proposed bill seems likely to pass in the near to mid-term.

- The absence of a comprehensive federal legal regime on AI has also led U.S. states to take their own actions. Many states have passed laws addressing deepfakes. Fewer have progressed from deepfakes to newer topics, such as banning algorithmic discrimination, as Colorado has. In California, the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act is a comprehensive bill that is currently pending. This bill aims to mitigate a wide range of potential risks posed by frontier models by imposing many compliance measures on developers of AI models

that present "hazardous capabilities." If adopted, California's regulators would become the de facto regulators of the AI industry in the U.S.

- Finally, it seems that U.S. authorities are primarily striving to address some, but not all, of the risks examined in Chapter 3. The federal government appears to be mainly focused on ensuring national and cyber security, especially for the most advanced AI systems, while still trying to locate an equilibrium between promoting innovation and preserving safety. States have ventured further by adopting some legislation targeting specific issues, such as deepfakes. As a result, some of the risks and challenges outlined in Chapter 3 remain unaddressed. The complex legal issues arising from the development of generative AI in copyright law will likely be left to the discretion of the courts. Privacy and data protection are not guaranteed due to the lack of a federal framework. The environmental impact of generative AI is not discussed. It appears that the strategy is to allow AI companies to evaluate the risks associated with their services and voluntarily manage their mitigation.

FIGURE 50. US initiatives and AI risks

Possible risks and challenges of generative AI	**Main provisions of the US legal framework**
Technical vulnerabilities (section 3.1.1)	• Voluntary commitments: leading AI companies committed to ensuring products are safe before introducing them to the public and to build systems that put security first. This includes: performing internal and external red teaming of models or systems, sharing information, investing in cybersecurity and insider threat safeguards, incentivize third-party discovery and reporting of issues and vulnerabilities (*section 5.3.2.B (2)*) • NIST has released various guidance (non-binding) documents focusing on risk management and cybersecurity. In particular, the NIST *AI Risk Management Framework* articulates various characteristics of trustworthy AI systems, including valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed (*section 5.3.2.B (3)(c)*). • Developers of dual use foundation models trained using computing power greater than 10^{26} FLOPS are subject to reporting obligations, especially about cybersecurity and safety measures (Executive Order 14110 s. 4.2(a)(i)) (*section 5.3.2.B (3)(e)*).
Factually incorrect content (section 3.1.2.)	• Voluntary commitments: leading AI companies committed to earn public's trust through developing and deploying mechanisms that enable users to understand if content is AI-generated, and publicly

	reporting model capabilities, limitations, and domains of appropriate and inappropriate us (*section 5.3.2.B (2)*) • The (non-binding) NIST *AI Risk Management Framework* articulates various characteristics of trustworthy AI systems, including validity and reliability (*section 5.3.2.B (3)(c)*).
Opacity (section 3.1.3)	• The (non-binding) NIST *AI Risk Management Framework* articulates various characteristics of trustworthy AI systems, including transparency, explainability and interpretability (*section 5.3.2.B (3)(c)*).
Misuse and abuse (section 3.2.1)	• Voluntary commitments: leading AI companies committed to publicly reporting domains of appropriate and inappropriate use (*section 5.3.2.B (2)*) • Entities that develop a large-scale computing cluster with theoretical maximum computing capacity of 10^{20} FLOPS for AI training, located in a datacenter with network connectivity >100 Gbit/s must report various information about their cluster (EO 14110 s. 4.2(a)(ii)) (*section 5.3.2.B (3)(e)*). • Executive Order 14110 directs the Secretary of Commerce to propose regulations that would require US providers of IaaS to report any transactions where foreign persons or entities use the infrastructure of IaaS companies to train large AI models that could be used for malicious cyber-enabled activity (EO 14110 s. 4.2(c)) (*section 5.3.2.B (3)(e)*). • NIST's paper entitled "Reducing Risks Posed by Synthetic Content" surveys existing technical standards, tools, methods, and practices for preventing generative AI from producing CSAM content or non-consensual deepfakes (*section 5.3.2.B(3)(c)*) • Some states have adopted statutes prohibiting sexual deepfakes, such as New York (NY SB 1042A), Texas (TX SB 1361), and Minnesota (MN HF 1370) (*section 5.3.3.A*)

Misinformation and Disinformation (section 3.2.2)	• Voluntary commitments: leading AI companies committed to developing and deploying mechanisms that enable users to understand if content is AI-generated, including provenance or watermarking (*section 5.3.2.B (2)*) • NIST's paper entitled "Reducing Risks Posed by Synthetic Content" surveys existing technical standards, tools, methods, and practices for authenticating content and tracking its provenance, labeling synthetic content (such as by watermarking) and detecting synthetic content (*section 5.3.2.B(3)(c))* • Various state laws mandate disclosure when AI is used to create content to influence an election, such as Florida (CS/HB 919 - AI Use in Political Advertising (2024)) or Wisconsin (2023 Act 123 (2024)) (*section 5.3.3.A)* • Tennessee prohibits the unauthorized use of deepfakes of any person's voice (ELVIS Act of 2024). (*section 5.3.3.A)*
Bias and discrimination (section 3.2.3)	• The (non-binding) NIST *AI Risk Management Framework* articulates various characteristics of trustworthy AI systems, including fairness with harmful bias managed (*section 5.3.2.B (3)(c)*). • Labor Department to publish guidance regarding nondiscrimination in hiring involving AI for federal contractors, and Federal Housing Finance Agency and CFPB encouraged to consider using their authorities to prevent discrimination (White House EO 14110 s. 7.3) (*section 5.3.2.A (2) and (3))* • Colorado requires developers and deployers of high-risk AI systems in consequential decision-making to use reasonable care to avoid algorithmic discrimination (SB24-205, Consumer Protections for Artificial Intelligence (2024)) (*section 5.3.3.A)*
New capabilities (section 3.2.5)	• Voluntary commitments: leading AI companies committed to sharing information among themselves and governments about dangerous or emergent capabilities (*section 5.3.2.B (2)*)
Open source models (section 3.2.6.A)	• White House EO 14110 – directed Commerce Department to solicit input on how to treat open source models (EO 14110 s. 4.6) (*section 5.3.2.B)*
Frontier Models (Section 3.2.6.B)	• Developers of dual use foundation models trained using computing power greater than 10^{26} FLOPS are subject to reporting obligations, especially about cybersecurity and safety measures (EO 14110 s. 4.2(a)(i)) (*section 5.3.2.B (3)(e))*).

Privacy and data protection (section 3.3.1)	• The (non-binding) NIST *AI Risk Management Framework* articulates various characteristics of trustworthy AI systems, including "privacy-enhanced" systems (section 5.3.2.B (3)(c)). • Col. Consumer Protections for AI Act cover profiling and automated decision making (*section 5.3.3.A)*
Copyrights (section 3.3.2)	• Pending litigation (*section 5.3.1.B)*
Impact on labor market (section 3.4.2)	• Reports commissioned by economic agencies and Labor Department to understand AI's impact on workers and how to support them in the event of displacement (EO 14110 s. 6) (*section 5.3.2.B (3))*
Environmental impact (section 3.4.3)	• Energy Department and science agencies directed to consider potential for AI to improve electric grid infrastructure and support development of AI tools (EO 14110 s. 5.2(g)) (*section 5.3.2.B (3))*

5.4. Ongoing regulatory initiatives

While the European Union, China, and the United States lead in developing AI policies and legislation, many other countries and regions worldwide have also initiated AI governance strategies. This section outlines the policies adopted by several countries, selected on the basis of recent developments, their presence in AI discussions, and the vitality of their AI ecosystems. However, this list is not exhaustive; numerous countries or regions with initiatives in this area are not included here.

The countries discussed in this section have adopted diverse strategies regarding artificial intelligence regulation. Some have chosen to implement comprehensive legislation, following the example set by the European Union. Others, which initially excluded legislative measures to prioritize innovation, are now reconsidering their stance and seriously contemplating the adoption of a binding framework in the near to medium term. Finally, some countries opt for non-binding ethical and technical guidelines.

5.4.1. Brazil

Brazil's effort to establish a legal framework for artificial intelligence has been influenced by the European Union's AI Act. Its work began in 2019 with the launch of the "Brazilian AI Strategy" (EBIA).[1844] The Strategy is a policy for developing Brazil's laws to promote responsible and ethical use of artificial intelligence in Brazil. In 2020, the Brazilian National Congress began consideration of Bill 21/2020, aimed at establishing the "Legal Framework of Artificial Intelligence." The bill was first introduced to the Chamber of Deputies,[1845] and was the first of what would eventually total four proposed bills.[1846]

After the Chamber of Deputies approved Bill 21/2020, the Federal Senate created a commission of experts in technology law and regulation. This Commission of Legal Jurists Responsible for Subsidizing the Drafting of an Alternative Bill on AI[1847] held a series of public hearings and international seminars to gather opinions from global experts. It also conducted six months of research and analysis into legislation and initiatives in other countries. By the end of the year, this Commission submitted a 900-page-long report, including a proposal for a new bill.[1848] On May 3, 2023, this draft, Bill 2338/2023, was introduced in the

[1844] Estratégia Brasileira de Inteligência Artificial (EBIA) [Brazilian Artificial Intelligence Strategy]: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/inteligencia-artificial (Braz.); Cristina Akemi Shimoda Uechi & Thiago Guimarães Moraes, *Brazil's path to responsible AI*, OECD (July 27, 2023), https://oecd.ai/en/wonk/brazils-path-to-responsible-ai.

[1845] Projeto de Lei 21/202: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2236340; *translated in* https://www.derechosdigitales.org/wp-content/uploads/Brazil-Bill-Law-of-No-21-of-2020-EN.pdf.

[1846] Before Bill 2338/2023 was introduced in the Federal Senate in May 2023, Bill 5.051/2019 and Bill 872/2021 were presented to the Chamber of Deputies, while Bill 21/2020 was introduced in the Federal Senate.

[1847] Comissão de Juristas responsável por subsidiar elaboração de substitutivo sobre inteligência artificial no Brasil (CJSUBIA): https://legis.senado.leg.br/comissoes/comissao?codcol=2504.

[1848] Senado Federal, Projeto de Lei N° 2338, de 2023 (Dispõe sobre o uso da Inteligência Artificial) [Bill 2338 of 2023 (Provides for the use of Artificial Intelligence)]: https://legis.senado.leg.br/sdleg-getter/documento?dm=9347622&ts=1715114415295&disposition=inline.

Brazilian Senate.[1849] It aimed to regulate the use of artificial intelligence, including algorithm design and technical standards.

In April 2024, the Senate announced that its Internal Temporary Commission on Artificial Intelligence published a new preliminary report with an updated proposal for Bill 2338/2023.[1850] A key innovation of this alternative version of the bill was the proposal of a dual oversight system, creating a National System for AI Regulation and Governance (SIA), coordinated by a competent authority to be appointed by the executive branch. On May 8, 2024, the Brazilian data protection authority (Autoridade Nacional de Proteção de Dados - ANPD) published its proposal for amendments to Bill 2338/2023.[1851] Specifically, the ANPD proposed the concept of a "general purpose AI system." It defined this as an AI system based on an AI model trained with large-scale databases, capable of an ample variety of different tasks and of serving different purposes including those for which it was not specifically designed or trained, and that may be used in different systems or applications.

5.4.1.A. A risk-based approach

In its most recent version, the Bill 2338/2023[1852] defines an AI system as:

"a computer system, with different degrees of autonomy, designed to infer how to achieve a given set of objectives, using machine learning and/or logic- and knowledge-based approaches, through machine and/or human-provider data, in order to produce predictions, recommendations, or decisions that may influence the virtual or real environment" (Chapter 1, Article 4, I).[1853]

Similar to the European Union's AI Act, the current bill adopts a risk-based approach, classifying AI systems in one of three risk categories:

1. excessive risk (for which use is prohibited entirely),
2. high-risk (which imposes a series of obligations on providers), and
3. non-high risk.

The criteria used to classify the risk level for an AI system are:

[1849] Projeto de Lei 2338, de 2023 (Dispõe sobre o uso da Inteligência Artificial) [Bill 2338 of 2023 (Provides for the use of Artificial Intelligence)], SENADO FEDERAL [FEDERAL SENATE]: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233 (Braz.); *translated in* https://mcusercontent.com/af97527c75cf28e5d17467eaa/files/248d109f-eeef-7496-4df1-12d29affb522/PL_23382023_Senado_ENG_VF.pdf.

[1850] Comissão Temporária Interna sobre Inteligência Artificial no Brasil [Internal Temporary Commission on Artificial Intelligence in Brazil], SENADO FEDERAL [FEDERAL SENATE] (24.04.2024): https://legis.senado.leg.br/comissoes/arquivos?ap=8139&codcol=2629.

[1851] ANPD apresenta propostas de alteração do substitutivo ao PL 2338, sobre inteligência artificial https://www.gov.br/anpd/pt-br/assuntos/noticias/anpd-apresenta-propostas-de-alteracao-do-substitutivo-ao-pl-2338-sobre-inteligencia-artificial.

[1852] Senado Federal, Projeto de Lei N° 2338, de 2023 (Dispõe sobre o uso da Inteligência Artificial) [Bill 2338 of 2023 (Provides for the use of Artificial Intelligence)]: https://legis.senado.leg.br/sdleg-getter/documento?dm=9347622&ts=1715114415295&disposition=inline.

[1853] *Id.*

- whether the system is implemented on a large scale;
- what the potential is for the system to have negative impact on the exercise of rights and freedoms;
- what the possibility is that the system will cause material or moral damage, irreversible damage, or discriminatory use; and
- whether the system negatively affects people from vulnerable groups, such as children, the elderly, or people with disabilities.

FIGURE 51 Classification of AI systems in the Brazilian Artificial Intelligence Bill

"Excessive risk" AI systems	**High-risk AI systems**
Article 13 prohibits the implementation and use of AI systems that: ● employ subliminal techniques to induce people to act in ways that could be harmful or dangerous to themselves or others; ● exploit any of the vulnerabilities related to a person's age, socio-economic situation, or physical or mental disability to induce harmful behaviors; and ● are used by public authorities to assess, classify, or rank individuals based on their social behavior or personality traits in an illegitimate or disproportionate manner.	High-risk AI systems (Article 15) are classified according to the above-mentioned criteria. These high-risk systems include: ● large-scale systems with extensive geographic or demographic reach and capable of adversely affecting individual rights and freedoms; ● systems with high potential to cause material or moral harm, as well as discrimination; ● systems that target vulnerable groups; ● systems that could have irreversible or difficult to reverse harmful outcomes, as well as those that have historical precedents of causing material or moral damage; ● systems with a low degree of transparency, explainability, and auditability, posing challenges for their control or oversight; ● systems that can significantly identify individuals or groups and that might compromise public health, safety, and the integrity of information; ● foundational, general-purpose, and generative AI models with systemic harmful potential, such as cybersecurity threats, integrity of electoral processes, and violence against vulnerable groups; ● biometric identification systems, except those used for authentication; and ● systems that could negatively impact informational integrity, the democratic process, and pluralism, for example, through the dissemination of disinformation and hate speech.

5.4.1.B. Text and data mining exception

Bill 2338/2023 seeks to create copyright exceptions for text and data mining processes used for developing AI systems.[1854] It specifies that such activities are permissible when conducted by research organizations, educational institutions, museums, archives, libraries, and journalistic entities, provided that:

- the copyrighted content was accessed in a legitimate manner;
- it is not used for commercial or business purposes;
- the activity does not primarily aim to reproduce, display, or disseminate the original work; and
- the use of copyrighted materials is necessary and proportional to the intended purpose, without unduly harming the economic interests of the copyright holders or competing with the expected use of the works.

This exception does not apply to profit-driven institutions operating or providing AI systems.

5.4.1.C. Obligations of AI systems providers and operators

Bill 2338/2023 would impose obligations on AI system providers[1855] and operators.[1856] All AI system providers must conduct a preliminary self-assessment analysis on their AI systems to determine the risk level before introducing the AI service to the market.[1857] AI system providers and operators must conduct algorithmic impact assessments when requested by the competent authority or whenever the AI system is deemed high-risk by the preliminary assessment. They must also report serious security incidents to the competent authority.[1858]

If a system is high-risk, its providers must fulfill an algorithmic impact assessment containing:

- known and foreseeable risks related to the AI system,
- benefits brought by the system, and
- information about risk mitigation measures, among other elements.

[1854] Projeto de Lei 2338, de 2023 (Dispõe sobre o uso da Inteligência Artificial) [Bill 2338 of 2023 (Provides for the use of Artificial Intelligence)], art. 54, SENADO FEDERAL [FEDERAL SENATE]: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233 (Braz.).

[1855] An AI system provider is "a natural or legal person, whether public or private, that develops an AI system, directly or by commission, with the intention of placing it on the market or applying it in a service provided by them, under their own name or brand, for consideration or free of charge." *Id.*, art. 4(II).

[1856] An AI system operator is "a natural or legal person, whether public or private, that employs or uses an AI system on their own behalf or for their benefit, unless the said system is used within the scope of a non-professional personal activity." *Id.*, art. 4(III).

[1857] *Id.*, art. 12 ("Before their introduction to the market or use in service, artificial intelligence agents must conduct a preliminary assessment of the artificial intelligence system, which will determine its risk level, based on the criteria provided in this chapter and sectoral best practices, according to the state of the art and technological development."); *Brazil: Senate Considers Bill Regulating AI*, DATAGUIDANCE (June 15, 2023), https://www.dataguidance.com/news/brazil-senate-considers-bill-regulating-ai.

[1858] Projeto de Lei 2338, de 2023 (Dispõe sobre o uso da Inteligência Artificial) [Bill 2338 of 2023 (Provides for the use of Artificial Intelligence)], art. 31, SENADO FEDERAL [FEDERAL SENATE]: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233 (Braz.).

The AI system provider must disclose their final assessment of their AI system to the competent authority, who shall organize them into a public database (with due regard to protect industrial and trade secrets).

All providers and operators of AI systems are required to implement governance policies and internal processes to mitigate risks, which includes transparency and security measures.[1859] In addition, providers and operators of high-risk AI systems must implement several key governance measures and internal processes, such as appointing a governance officer, maintaining adequate technical documentation, and using log registers, reliability tests, measures to mitigate discriminatory bias, and measures to ensure explainability and transparency.[1860]

5.4.1.D. User rights and liability

The bill provides for certain specific rights for individuals who are affected by AI systems. These include the right to:

- information about their interactions with an AI system before they use it;
- an explanation about decisions, recommendations, or predictions made by AI systems;
- correct incomplete, inaccurate, or outdated data used by AI systems;
- nondiscrimination and to correction of discriminatory biases; and
- privacy and protection of personal data.

The bill establishes comprehensive liability rules, holding providers and users accountable for damages caused by AI systems, whether the harm is material, moral, individual, or collective, and irrespective of the AI system's level of autonomy. Article 32 categorizes AI systems according to their risk level. Providers and operators of high-risk or excessive-risk AI systems are held strictly liable for any damages. For other AI systems, fault is presumed, shifting the burden of proof in favor of the victim, thus facilitating damage claims. The bill also specifies various exemptions from liability, such as a case where the harm is caused solely by the victim, a third party, or an external force majeure event.

5.4.1.E. Sandboxes

Finally, the Brazilian AI bill seeks to promote technological innovation by regulating testing environments, known as "sandboxes." These controlled settings allow new AI technologies to be tested under regulatory supervision before broader deployment. This approach permits real-world testing of AI systems without exposing them to the full extent of legal and regulatory consequences. It provides developers and regulators with valuable insights into the systems' operations and potential impacts. The sandbox is designed to allow temporary relaxation of certain regulations, upon request by the AI system provider and authorization by the competent authority. Despite this relaxation, the systems remain under scrutiny to ensure they meet safety, efficacy, and compliance standards with overarching legal requirements.

[1859] *Id.*, art. 19.
[1860] *Id.*, art. 19-20.

5.4.1.F. Enforcement

Based on the current draft, Bill 2338/2023 will apply to the development, implementation, and use of AI systems within Brazilian territory, without distinguishing between national and foreign entities.

The bill requires the government to assign a supervisory authority to a public body to implement, oversee, and enforce the provisions of the future law. The recent report by the Senate's Temporary Commission on AI provides for the implementation of a National System of Artificial Intelligence Regulation and Governance (SIA).[1861] This SIA is to be a mechanism, coordinated by the competent authority (yet to be determined), to supervise and guide the use of AI in cooperation with other agencies and regulatory bodies. It is intended to ensure the comprehensive implementation and enforcement of the future AI law across Brazil. It will be responsible for regulating high-risk AI systems, ensuring that the potential for technological progress is balanced against the risk. The SIA's role would not be static; it would involve continuous monitoring and reclassification of AI systems to reflect evolving risks and technological advancements.

The Senate report suggests that this authority could be the existing data protection authority, the ANPD, which would need to be enhanced and expanded for this purpose. This system would also involve existing regulatory agencies, such as the Brazilian Health Regulatory Agency (Anvisa)[1862] and the National Agency for Telecommunications (Anatel),[1863] which wouldoversee AI in their respective areas.

Conclusion

If approved by the Federal Senate's temporary commission on AI, Bill 2338/2023 will proceed to the full Senate and then to the Chamber of Deputies. This law, particularly following recent amendments to its original text, is anticipated to have strong prospects for adoption in Brazil. The Senate is expected to vote on it in the coming months. The final version of the bill is expected to closely align with the preliminary report recently released by the Senate's temporary AI commission. Should this occur, the Brazilian AI bill would represent a pivotal development in creating a robust framework for the regulation of artificial intelligence within the nation. By integrating international best practices and tailoring them to Brazil's specific needs, the bill strives to foster technological innovation while ensuring the protection of individual rights and the upholding of societal values.

5.4.2. Canada

[1861] Comissão Temporária Interna sobre Inteligência Artificial no Brasil [Internal Temporary Commission on Artificial Intelligence in Brazil], SENADO FEDERAL [FEDERAL SENATE] (24.04.2024): https://legis.senado.leg.br/comissoes/arquivos?ap=8139&codcol=2629 (Braz.).

[1862] Agência Nacional de Vigilância Sanitária (Anvisa): https://antigo.anvisa.gov.br/en/english.

[1863] Agência Nacional de Telecomunicações: https://www.gov.br/anatel/pt-br.

As a country with a long and established history of foundational AI research,[1864] Canada is still in the process of adopting its first comprehensive AI regulation. The bill is the Artificial Intelligence and Data Act (AIDA), introduced as part of Bill C-27, the Digital Charter Implementation Act of 2022. This proposed statute aims to ensure the responsible development and use of AI systems in Canada. Meanwhile, Canada has introduced a Voluntary Code of Conduct for the responsible development and management of advanced generative AI systems. This voluntary code serves as an interim measure while Canada finalizes its legislative framework.

5.4.2.A. The Digital Charter Implementation Act (Bill C-27)

Canada has had national privacy protection laws since 2000. The main personal data protection framework is the Personal Information Protection and Electronic Documents Act (PIPEDA). This law established a consent-based privacy regime where commercial entities must obtain individuals' consent before collecting, using, or disclosing their information.[1865] When the EU's General Data Protection Regulation (GDPR) came into effect, policy experts believed that Canada's PIPEDA likewise needed an update.[1866] To that end, the Canadian Parliament drafted the Digital Charter Implementation Act of 2022.

The Digital Charter Implementation Act (also known as Bill C-27)[1867] was introduced in June 2022 and represents the country's most salient attempt yet to modernize its rule of law to address both new privacy risks and AI. Bill C-27 seeks to replace part one of PIPEDA, which protects personal information within the private sector.[1868] It also seeks to impose regulations on commercial entities developing and deploying AI systems. Altogether, Bill C-27 has three sections:

1. The Consumer Privacy Protection Act (CPPA) would establish new rules for the commercial use of personal information and would strengthen the Privacy Commissioner's enforcement role.
2. The Personal Information and Data Protection Tribunal Act (PIDPTA) would establish a "Privacy Tribunal" that would settle disputes and enforce new administrative penalties.
3. The Artificial Intelligence and Data Act (AIDA) would address commercial AI systems and impose transparency, nondiscrimination, and safety measures for "high-impact" AI systems. AIDA is the only part of Bill C-27 to deal directly with artificial intelligence and

---

[1864] *How Canada's Unique Research Culture Has Aided Artificial Intelligence*, THE ECONOMIST (November 4, 2017), https://www.economist.com/the-americas/2017/11/04/how-canadas-unique-research-culture-has-aided-artificial-intelligence.

[1865] PIPEDA Requirements in Brief, OFFICE OF THE PRIVACY COMMISSIONER OF CANADA (May 2019), https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/#_h3.

[1866] Ryan Martin, *Canada's Federal Government proposes Changes to Privacy Act*, OGLETREE DEAKINS (July 27, 2022), https://ogletree.com/insights-resources/blog-posts/canadas-federal-government-proposes-changes-to-privacy-act.

[1867] Bill C-27, Bill 44-1 (2021), https://www.parl.ca/LegisInfo/en/bill/44-1/C-27.

[1868] Personal Information Protection and Electronic Documents Act (S.C. 2000, c. 5) (Canada), https://laws-lois.justice.gc.ca/eng/acts/p-8.6/.

has received the most attention from legislators and the media.[1869] Where the other two sections of Bill C-27 build on previous rules under the PIPEDA, AIDA introduces a *new* framework.

After successfully passing its second reading, Bill C-27 was referred to the Standing Committee on Industry, Science, and Technology of the House of Commons on April 24, 2023.[1870] Subsequently, on November 28, 2023, the Minister of Innovation, Science, and Industry (ISI) presented the Committee with substantial amendments regarding the AIDA section of Bill C-27.[1871] The proposed amendments indicate an intent to harmonize the legislative text of Canada's bill with the European Commission's proposals, OECD principles, and the U.S. NIST framework. The amendments have not yet been officially adopted, and it remains uncertain whether the overall Bill C-27 will see a vote anytime before Canada's 2025 federal election.

AIDA seeks to impose varying obligations based on the type of AI system involved (i.e., general-purpose, machine-learning models, or high-impact systems) and the position of the different AI players within the AI value chain.

1) Definition of AI systems

AIDA, as the government seeks to amend it, would include the following definitions:

- **"**Artificial intelligence system**"** means a "system that, using a model, makes inferences in order to generate output, including predictions, recommendations or decisions."[1872]
- "General-purpose system" means an "artificial intelligence system that is designed for use, or that is designed to be adapted for use, in many fields and for many purposes and activities, including fields, purposes and activities not contemplated during the system's development."[1873]

[1869] Alex LaCasse, *Canadian Parliament's Bill C-27 hearing delves deeper into AIDA*, IAPP (December 8, 2023), https://iapp.org/news/a/canadian-parliaments-bill-c-27-hearing-delves-deeper-into-aida/; *Canadian Public Employees' Union Comes Out Against AIDA*, IAPP (February 23, 2024), https://iapp.org/news/a/canadian-public-employees-union-comes-out-against-aida/; *Artificial Governance: AIDA Repeats the Failed Patterns of Digital Regulatio*n , CENTRE FOR INTERNATIONAL GOVERNANCE INNOVATION (December 18, 2023), https://www.cigionline.org/articles/artificial-governance-aida-repeats-the-failed-patterns-of-digital-regulation/

[1870] *Id.*

[1871] *Remarks by the Honourable Joel Lightbound, M.P. to the Standing Committee on Industry, Science, and Technology*, CANADA, MINISTER OF INNOVATION, SCIENCE, AND INDUSTRY (November 28, 2023), https://www.ourcommons.ca/content/Committee/441/INDU/WebDoc/WD12751351/12751351/MinisterOfInnovationScienceAndIndustry-2023-11-28-Combined-e.pdf.

[1872] Bill C-27, cl. 39 (44th Parliament, 1st session, November 22, 2021, to present).

[1873] *Id.*

- "Machine-learning model" means a "digital representation of patterns identified in data through the automated processing of the data using an algorithm designed to enable the recognition or replication of those patterns."[1874]

The proposed amendments clarify that an AI system may be a general-purpose system and a high-risk system at the same time. They also clarify the meaning of a "high impact system," a concept that previous drafts of the bill had left undefined. "High-impact systems" would be defined by their potential applications within specific categories, some of which pertain to particular sectors.[1875] An AI system would be classified as a "high-impact system" if its intended use falls within one of several categories outlined in the proposed amendments. Specifically, an AI system will be deemed "high-impact" if it is employed in: making employment-related decisions, deciding whether to provide government services to an individual, processing biometric information, moderating online content, and any uses within the context of healthcare and emergency services, judicial processes, and law enforcement.[1876]

2) Obligations related to AI systems

AIDA generally imposes obligations on AI companies to ensure human oversight of AI technologies and to implement standards of transparency, fairness, equity, safety, and accountability. The amendments seek to delineate the roles and responsibilities of various actors within the AI value chain, in accordance with the AIDA Companion Document.[1877] This includes implementing data governance measures, conducting an impact assessment, and ensuring "appropriate human oversight."[1878]

Specific obligations are tailored to each actor's role within the AI value chain.[1879] For general-purpose systems and high-impact AI systems, AIDA targets individuals or organizations responsible for managing these systems, making them available, and introducing them for the first time in cross-border trade. For machine-learning models, the bill targets individuals or organizations first making a machine-learning model available for incorporation into a high-impact system in cross-border trade.

If an AI system is categorized as "high impact," the system's designers, developers, creators, and managers face specific obligations related to identifying, assessing, tracking, and mitigating the system's potential

---

[1874] *Id.*

[1875] *Artificial Intelligence and Data Act (AIDA) Companion Document*, CANADIAN MINISTRY OF INNOVATION, SCIENCE AND ECONOMIC DEVELOPMENT, https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document#s6.

[1876] Remarks by the Honourable Joel Lightbound, M.P. ti the Standing Committee on Industry, Science, and Technology, Canada, MINISTER OF INNOVATION, SCIENCE, AND INNOVATION (November 28, 2023), https://www.ourcommons.ca/content/Committee/441/INDU/WebDoc/WD12751351/12751351/MinisterOfInnovationScienceAndIndustry-2023-11-28-Combined-e.pdf.

[1877] *Artificial Intelligence and Data Act (AIDA) Companion Document*, *supra* note 1875.

[1878] *Id.*

[1879] *Id.*, §7-12.

for causing harm.[1880] Entities that first make a machine-learning model available for incorporation into a high-impact system (i.e., developers) must establish measures to identify, assess, and mitigate the risks of biased output before the model is released. They must also ensure that data used in developing the model comply with regulations. Those who first make high-impact systems available must ensure that risk mitigation measures are in place and have been tested. Managers of high-impact systems are responsible for conducting tests to verify the effectiveness of these mitigation measures. Operators of high-impact systems must also establish mechanisms for users to provide feedback on the system's performance.

The proposed amendments to AIDA require those who make a general-purpose system or high-impact system available, or who manage their operations, to establish and maintain written accountability frameworks.[1881] These frameworks must include a description of the roles, responsibilities, and reporting structure for all personnel involved, policies and procedures for managing risks related to the system and its data, and any other requirements specified by regulation.

3) Enforcement

The initial draft of the AIDA section of Bill C-27 granted substantial authority for its implementation to the Minister of Innovation, Science, and Economic Development (ISED). Within that department, the Minister of Innovation, Science, and Industry (ISI) would appoint a senior official from their program to serve as the Artificial Intelligence and Data Commissioner. This Commissioner would assist the Minister in the administration and enforcement of AIDA. The proposed amendments reallocate certain powers from the Minister to the Commissioner. Under the proposed amendments, the Commissioner would have central responsibility to enforce AIDA.

Under the proposed amendments, the Commissioner would also have the authority to compel from those who make a general-purpose or high-impact system available, or manage their operations, an accountability framework. They would be able to offer guidance or recommendations for necessary corrective measures.[1882] They could require entities who make available or manage any AI system to provide an assessment of whether the AI system is subject to AIDA.

If the Commissioner has reasonable grounds to believe that a person has contravened or is likely to contravene certain sections of AIDA, the Commissioner could conduct an audit, mandate any person to conduct an audit, or require the engagement of an independent auditor. Additionally, they would have the power to communicate information to and from other regulators, including the Privacy Commissioner, the Canadian Human Rights Commission, the Commissioner of Competition, and the Financial Consumer Agency of Canada.

[1880] Alan Macek, et al. *Canada outlines proposed regulation of AI systems in companion paper to the Artificial Intelligence and Data Act*, DLA PIPER (April 18, 2023), https://www.dlapiper.com/en/insights/publications/2023/04/canada-releases-companion-paper-on-artificial-intelligence-and-data-act.

[1881] *Artificial Intelligence and Data Act (AIDA) Companion Document*, *supra* note 1875, §12.

[1882] *Id.*, §13-15.

The ISI Minister would have the authority to order an individual or organization to take actions to achieve compliance with the law.[1883] The Minister could mandate the cessation of availability or termination of operation of a system if compliance is deemed impossible[1884] or if there are reasonable grounds to believe that the use of the system "gives rise to a risk of imminent and serious harm."[1885]

AIDA provides for the imposition of monetary penalties for violations, but it does not specify who will set and enforce these penalties. Contravention of Sections 6 through 12 and certain other offenses are subject to the following penalties:[1886]

-For conviction on indictment, companies would face a fine of up to the greater of CAD 10 million or 3% of the company's gross global revenues in the preceding financial year; individuals would be subject to a fine at the discretion of the court.

-For summary conviction, companies could be fined up to the greater of CAD 5 million or 2% of their gross global revenues in the preceding financial year, and individuals may be fined up to CAD 50,000.

Under Sections 38 and 39 of AIDA, general offenses carry a strict penalty for such things as unlawful possession or utilization of personal information or designing, developing, using, or making available an AI system that could severely harm an individual:[1887]

-For conviction on indictment, companies could be fined up to the greater of CAD 25 million or 5% of their gross global revenues in the preceding financial year; individuals could be fined at the discretion of the court, sentenced to a term of imprisonment of up to five years less a day, or both.

-For summary conviction, companies would be fined up to the greater of CAD 20 million or 4% of their gross global revenues in the preceding financial year; individuals could be fined up to CAD 100,000, sentenced to a term of imprisonment of up to two years less a day, or both.

4) Reception

[1883] *Id.*, §16.
[1884] *Id.*, §16(b).
[1885] *Id.*, §17(1).
[1886] *Id.*, §30(3).
[1887] Under Sections 38 and 39, general offenses are committed if a person:
(i) "for the purpose of designing, developing, using or making available for use an AI system, possesses – within the meaning of subsection 4(3) of the Criminal Code – or uses personal information, knowing or believing that the information is obtained or derived, directly or indirectly, as a result of (a) the commission in Canada of an offence under an Act of Parliament or a provincial legislature; or (b) the an act or omission anywhere that, if it had occurred in Canada, would have constituted such an offence;" or
(ii) "(a) without lawful excuse and knowing that or being reckless as to whether the use of an AI system is likely to cause serious physical or psychological harm to an individual or substantial damage to an individual's property, makes the AI system available for use and the use causes such harm or damage; or (b) with intent to defraud the public and to cause substantial economic loss to an individual, makes an AI system available for use and its use causes that loss."

From its introduction, AIDA has been debated by legal experts,[1888] policymakers,[1889] and civil society organizations.[1890] Members from each community have been vocal in their concerns. Among their worries, in the beginning, was the legislation's lack of clarity and, in some cases, deferment of key definitions and regulatory procedures.[1891] For instance, while the law is intended to mitigate the harms stemming from "high-impact systems,"[1892] there was disclarity in the bill's first draft over what constitutes such a system. The amendments issued by the ISI Minister in November 2023 sought to update and clarify the definition of a "high-impact system." To that end, the proposed amendments included a schedule with seven defined "high-impact" classes to inform whether an AI system qualifies as "high impact."[1893]

Still, many of the same critics point to concerns over the law's scope and stakeholder engagement. AIDA targets commercial enterprises but, notably, does not cover AI use in the public sector.[1894] The unaccounted for government use of AI worries some who argue that it should be addressed.[1895] In addition, when the government introduced AIDA in June 2022, it did so without engaging in the kind of public consultation process that some argue such impactful legislation warrants.[1896] One preliminary analysis of Bill C-27's stakeholder engagement found that, of the 253 stakeholders engaged,[1897] 216 belonged to the business sector and only a few were "representatives of those who might be affected, and those at risk."[1898] These and other reported shortcomings led a number of civil society organizations and experts to submit a letter to the ISED Ministry calling for AIDA's removal from the three-part Bill C-27 and to, instead, reintroduce a

---

[1888] Yves Faguy, *Canada Mustn't Rush into Legislating AI*, NATIONAL MAGAZINE (Nov. 17, 2023), https://nationalmagazine.ca/en-ca/articles/law/hot-topics-in-law/2023/canada-mustn-t-rush-into-legislating-ai.

[1889] Joanna Redden, *Federal Government's Proposed AI Legislation Misses the Mark on Protecting Canadians*, WESTERN NEWS (April 12, 2024), https://news.westernu.ca/2024/04/proposed-ai-legislation/.

[1890] Howard Solomon, *Experts Urge Changes to Proposed Canadian Privacy, AI Laws Before Today's Hearing*, FINANCIAL POST (September 26, 2023), https://financialpost.com/technology/experts-urge-changes-to-proposed-canadian-privacy-ai-laws-before-todays-hearing.

[1891] Faguy, *supra* note 1888; Carolyn Gruske, *Critics Say Artificial Intelligence and Data Act Needs to Focus More on Rights, Not Just Business*, CANADIAN LAWYER MAG., https://www.canadianlawyermag.com/practice-areas/privacy-and-data/critics-say-artificial-intelligence-and-data-act-needs-to-focus-more-on-rights-not-just-business/380552.

[1892] *Id.*, §5,7.

[1893] Canada, Parliament, Standing Committee on Industry and Technology, Appearance of the Hon. François-Philippe Champagne, Minister of Innovation, Science and Industry (Nov. 28, 2023), https://www.ourcommons.ca/content/Committee/441/INDU/WebDoc/WD12751351/12751351/MinisterOfInnovationScienceAndIndustry-2023-11-28-Combined-e.pdf.

[1894] Gruske, *supra* note 1891.

[1895] *AIDA -- Priority Recommendations Package*, OPENMEDIA, https://openmedia.org/assets/AIDA_--_Priority_Recommendations_Package_-_FINAL.pdf.

[1896] *Key Stakeholders Call For Withdrawal of Controversial AI Legislation*, PACC, https://pacc-ccap.ca/aida-open-letter/; *Joint Call for AIDA to Be Sent Back for Meaningful Public Consultation and Redrafting*, CENTRE FOR FREE EXPRESSION, https://cfe.torontomu.ca/page/joint-call-aida-be-sent-back-meaningful-public-consultation-and-redrafting.

[1897] Andrew Clement, *Preliminary Analysis of ISED's C-27 List of 300 Stakeholder Consultation Meetings*, (Dec. 6, 2023), https://ssrn.com/abstract=4658004 or http://dx.doi.org/10.2139/ssrn.4658004.

[1898] *Id.*

"significantly improved" AIDA.[1899] Such calls were renewed in December 2023[1900] and April 2024.[1901] Currently, AIDA is sitting under committee consideration in the House of Commons with an undetermined enactment timeline.[1902] As the legislative process unfolds, debates and amendments are expected to continue to shape the final form of the bill. Moreover, international regulatory trends, particularly from Europe, are anticipated to influence future amendments and the enactment of regulations following the law's implementation.

5.4.2.B The Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems

In conjunction with its own efforts on Bill C-27, the ISED Ministry issued a non-binding code of conduct for firms that develop or manage generative AI with general-purpose capabilities.[1903] The document, titled "Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems," was released in September 2023, a month prior to the release of the G7's International *Code of Conduct* (*see below section 6.3*). This code aims to establish best practices and set standards to foster public trust and ensure ethical deployment of AI.

Similar to other codes, Canada's code of conduct enumerates six core principles (accountability, safety, fairness and equity, transparency, human oversight and monitoring, and validity and robustness) that apply to developers and managers, albeit with differing responsibilities for each.[1904] For instance, to ensure "transparency," it is incumbent on a manager, but not a developer, to make sure that an AI system's dialogue that might be confused for that of a human is clearly marked as coming from an AI system.[1905] On the other hand, it is incumbent on developers, but not managers, to assess training data for quality and mitigate harmful biases in order to ensure "fairness and equity."[1906] In total, 23 organizations have committed to the voluntary code of conduct as of April 2024. They include AI firms, such as IBM, Bluedot, and Cohere, and other organizations and research institutes, such as the Council of Canadian Innovators and the Vector Institute.

[1899] *AIDA Joint Letter for Sign On*, INTERNATIONAL CIVIL LIBERTIES MONITORING GROUP (Sept. 25, 2023), https://iclmg.ca/wp-content/uploads/2023/09/AIDA-JOINT-LETTER-FOR-SIGN-ON.pdf.

[1900] *Final Draft - AIDA Committee Split Letter*, OPENMEDIA (December 14, 2023), https://openmedia.org/assets/AIDA_Civil_Society_Letter_to_INDU_HoC_Committee.pdf.

[1901] AIDA Joint Letter, OPENMEDIA (April 24, 2024), https://openmedia.org/assets/AIDA_joint_letter.pdf.

[1902] Michael M. Gallagher, *Canada's Artificial Intelligence and Data Act (AIDA) 2024: A Comprehensive Guide*, COX & PALMER (April 11, 2024), https://coxandpalmerlaw.com/publication/aida-2024/.

[1903] Voluntary Code of Conduct for Responsible Development and Management of Advanced Generative AI Systems, GOVERNMENT OF CANADA (September 2023), https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems.

[1904] *Id.*

[1905] *Id.*

[1906] *Id.*

The code of conduct is seen as a stopgap measure for what the ISED Ministry hopes is future, binding AI legislation. In its news release for the code of conduct, the Ministry noted that the code was intended to act as a "critical bridge between now and when that legislation [AIDA] would be coming into force."[1907]

Conclusion

Currently, no specific legislation regulating AI exists in Canada. The Canadian government took its initial steps toward regulating artificial intelligence with the introduction of Bill C-27, the Digital Charter Implementation Act of 2022. This bill includes a specific section, known as the Artificial Intelligence and Data Act (AIDA), which presents the foundational text for Canada's first law designed to oversee the development and deployment of AI systems. Bill C-27 also includes the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act, marking Canada's second attempt to reform its privacy laws. In its proposed amendments to AIDA, the Canadian government aims to harmonize the Act with other regulatory frameworks, including the EU Artificial Intelligence Act. This alignment seeks to ensure that Canada's legislation is interoperable and consistent with international best practices.

Additionally, in September 2023, the federal government announced the Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, to which 23 organizations have committed. This Code seeks to provide Canadian companies with interim common standards, allowing them to voluntarily demonstrate responsible development and use of generative AI systems until formal regulations are enacted.

5.4.3. India

India is one of the prominent digital economies in Asia and has been witnessing rapid growth in the adoption of digital services over the past few years. India has also prioritized the development and adoption of artificial intelligence in its policy initiatives for the coming years. In March 2024, India announced an allocation of over $1.25 billion for the India AI Mission, which will cover various aspects of AI, including computing infrastructure capacity, skilling, innovation, datasets, and safe and trusted AI.[1908]

The Indian government is also cognizant of the possible harms and risks arising from technology. After years of deliberation, it recently enacted a national data protection law. Now, India is formulating regulatory frameworks to address AI-related risks and concerns.

5.4.3.A Existing legal frameworks

[1907] *Minister Champagne Launches Voluntary Code of Conduct Relating to Advanced Generative AI Systems*, CANADIAN MINISTRY OF INNOVATION, SCIENCE AND ECONOMIC DEVELOPMENT (Sept. 27, 2023), https://www.canada.ca/en/innovation-science-economic-development/news/2023/09/minister-champagne-launches-voluntary-code-of-conduct-relating-to-advanced-generative-ai-systems.html#.

[1908] *Cabinet Approves Over Rs 10,300 Crore for IndiaAI Mission, will Empower AI Startups and Expand Compute Infrastructure Access*, PRESS INFO. BUREAU (Mar. 7, 2024), https://pib.gov.in/PressReleasePage.aspx?PRID=2012375.

India does not currently have any legislation or legislative proposals that directly address AI. However, there are certain legislative and policy instruments that could have an impact on AI-related applications or that have sought to address regulatory and policy concerns arising from AI.

1) Developments under Indian copyright law

India's laws related to intellectual property do not explicitly address AI-related issues. The Digital News Publisher Association (DNPA), a major industry body representing 17 top media publishers in India, wrote to the government in January 2024, expressing concerns about copyright violations by AI models and their uncompensated use of content from news publishers as training data for generative AI models.[1909]

In response to a parliamentary question in February 2024, India's Minister for Commerce and Industry stated that India's current legal framework for intellectual property is sufficient to protect AI-generated works and that there is no proposal to amend the law in the context of AI-generated content. He also stated that the exclusive economic rights of a copyright owner under the Copyright Act of 1957 require generative AI service providers to obtain permission from copyright owners to use their works for commercial purposes, if such use is not covered under the "fair dealing" exceptions.[1910] While the minister's remarks may reflect the Indian government's official stance, responses to parliamentary questions are not legally enforceable. However, at least one major industry body representing India's music recording industry welcomed the government's clarification.[1911]

Finally, in an interview in April 2024, India's Minister of Electronics and Information Technology stated that the Indian government is working on a new AI law to protect the interests of news publishers and content creators, in addition to addressing user harms.[1912]

2) The Digital Personal Data Protection Act of India

An Indian law that will likely affect generative AI development is the Digital Personal Data Protection (DPDP) Act of India, which officially became a law in August 2023. The DPDP Act encompasses a wide range of applications, adopting the methodology of the EU's General Data Protection Regulation (GDPR) in defining "personal data" and extending its reach to all entities processing personal data, irrespective of their size or private status.[1913] The DPDP Act gives to individuals whose personal data have been collected

---

[1909] Annapurna Roy, *Indian publishers seek rules for copyright protection against generative AI models*, ECON. TIMES (Jan. 26, 2024), https://economictimes.indiatimes.com/tech/technology/indian-publishers-seek-rules-for-copyright-protection-against-generative-ai-models/articleshow/107154425.cms?from=mdr.

[1910] *Existing IPR regime well-equipped to protect AI generated works, no need to create separate category of rights*, PRESS INFO. BUREAU (Feb. 9, 2024), https://pib.gov.in/PressReleasePage.aspx?PRID=2004715.

[1911] Blaise Fernandes, LINKEDIN, https://www.linkedin.com/posts/blaise-fernandes-297b6612_existing-ipr-regime-well-equipped-to-protect-activity-7162020331333795840-vHgG/ (last visited May 3, 2024).

[1912] Surabhi Agarwal & Yash Aryan, *New AI law to secure rights of news publishers: Ashwini Vaishnaw*, ECON. TIMES (Apr. 05, 2024), https://economictimes.indiatimes.com/tech/technology/exclusive-new-ai-law-to-secure-rights-of-news-publishers-ashwini-vaishnaw/articleshow/109043916.cms?from=mdr.

[1913] Raktima Roy & Gabriela Zanfir-Fortuna, *The Digital Personal Data Protection Act of India, Explained*, FUTURE OF PRIVACY F. (Aug. 15, 2023), https://fpf.org/blog/the-digital-personal-data-protection-act-of-india-explained/.

and used the rights to notice, access, and erasure. It also requires data fiduciaries and data controllers to erase the collected personal data of users once the primary purpose of their collection is met. Furthermore, the law would establish the Data Protection Board of India, which would be empowered to investigate complaints and levy fines. Compared to GDPR and its contemporaries, this law includes wide exemptions for government actors and no heightened protection for special categories of data.[1914]

With regards to generative AI, some provisions of the DPDP Act appear to be aimed at protecting the ability to train AI models with personal data. For instance, the DPDP Act does not apply to personal data that are publicly available, provided that the data were made public by the individual to whom they pertain.[1915] The DPDP Act also does not apply to the processing of personal data necessary for research, archiving, or statistical purposes, subject to certain conditions being met. However, the DPDP Act would still apply if the processing done for research, archiving, or statistical purposes was used to make decisions related to the person whose data was processed.[1916]

3) Litigation related to deepfakes

There have been several court cases in India addressing issues related to the creation of deepfakes. A well-known movie actor filed a case against multiple websites, alleging that they were misusing his personality rights.[1917] Some of the examples of misuse included different kinds of deepfakes of the actor generated through the websites, such as using the actor's face on Disney characters. A state-level court found that commercial misuse of a celebrity's personality elements —image, voice, or likeness— violated the actor's fundamental right to privacy under the Constitution of India. The court specifically noted that "technological tools that are now freely available make it possible for any illegal and unauthorized user to use, produce or imitate any celebrity's persona, by using any tools including Artificial Intelligence." The court consequently granted an injunction restraining the defendants from using "the name, likeness, image, voice, personality or any other aspects of [the actor's] persona to…misuse the said attributes using technological tools such as Artificial Intelligence, Machine Learning, deepfakes, face morphing, GIFs either for monetary gains or otherwise."[1918]

Meanwhile, a public interest litigation case is currently pending before a state-level court, where the petitioner asks the court to direct the federal government to identify and block websites providing access to deepfake technology and to formulate guidelines to develop a mechanism and framework for the regulation of AI.[1919]

[1914] *Id.*

[1915] Digital Personal Data Protection (DPDP) Act, 2023, § 3(c)(ii).

[1916] *Id.*, § 17(2)(b).

[1917] Anil Kapoor v. Simply Life India & Ors., Delhi High Court, CS(COMM) 652/2023, https://dhcappl.nic.in/dhcorderportal/GetQROrder.do?ID=pms//2023//10001882169537605978 2_77267_2023.pdf.

[1918] *Id.*

[1919] Chaitanya Rohilla v. Union of India, Delhi High Court, W.P.(C) 15596/2023, https://dhcappl.nic.in/dhcorderportal/GetQROrder.do?ID=mmh//2023//10001856170178545500 0_27500_155962023.pdf.

4) Advisories issued by India's IT ministry

India's Ministry of Electronics and Information Technology (MEITY) issued advisories in 2023 related to deepfakes and generative AI. While there is uncertainty about the legally binding nature of these advisories, they give a sense of how seriously the Indian government is concerning itself with generative AI and how such concerns could be addressed in future regulations.

In November and December 2023, MEITY issued two advisories, requiring online platforms to take steps to tackle deepfakes and other disinformation-related content.[1920] These advisories reiterated existing obligations under Indian law for steps the online platforms should take to identify misinformation, prevent users from uploading any misinformation content, and expeditiously act on such content within 36 hours.[1921] These advisories essentially extended these existing obligations to apply to deepfakes. MEITY also posted a message on Twitter (rebranded as X in July 2023), saying that the December 2023 advisory was issued after holding two consultations, known as "Digital India Dialogues," specifically on the issue of tackling deepfakes with the relevant stakeholders.[1922]

MEITY also issued an advisory to online platforms, advising them to comply with various requirements for moderating online content in order to continue enjoying an immunity from liability (safe harbor protection) under Indian law. MEITY issued the advisory after India's IT minister shared a post on X/Twitter about an output from Google's Gemini that stated India's current prime minister has been accused of implementing "fascist" policies. The IT minister posted a message on X/Twitter stating that such outputs violate existing Indian law.[1923] Google reportedly took remedial measures to address this issue.[1924]

Following this incident, MEITY issued an advisory on March 1, 2024, requiring any "unreliable Artificial Intelligence model(s) /LLM/Generative AI, software(s) or algorithm(s)" to seek the "explicit permission" of the Indian government before deploying such models for "users on the Indian Internet." Companies were also directed to submit compliance status reports by March 15, 2024.[1925]

[1920] *Union Government issues advisory to social media intermediaries to identify misinformation and deepfakes*, PRESS INFO. BUREAU (Nov. 7, 2023), https://pib.gov.in/PressReleasePage.aspx?PRID=1975445; *MEITY issues advisory to all intermediaries to comply with existing IT rules*, PRESS INFO. BUREAU (Dec. 26, 2023), https://pib.gov.in/PressReleseDetailm.aspx?PRID=1990542.

[1921] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.

[1922] Rajeev Chandrasekhar (@Rajeev_GoI), TWITTER (Dec. 26, 2023), https://twitter.com/Rajeev_GoI/status/1739665244992291068.

[1923] Rajeev Chandrasekhar (@Rajeev_GoI), TWITTER (Feb. 22, 2024), https://x.com/Rajeev_GoI/status/1760910808773710038.

[1924] Soumyarendra Barik, *After Centre's threat to send notice, Google says 'addressed' AI response on PM Modi*, INDIAN EXPRESS (Feb. 24, 2024), https://indianexpress.com/article/india/govt-to-send-notice-to-google-over-its-ai-reply-to-query-on-pm-modi-9177978/lite/.

[1925] Aditi Agrawal, *Under-testing AI models must get govt permission before deployment: MEITY*, HINDUSTAN TIMES (March 2, 2024,), https://www.hindustantimes.com/india-news/undertesting-ai-models-must-get-govt-permission-before-deployment-meity-101709390335142.html.

This advisory faced criticism. In addition to questioning its legal basis, various stakeholders argued that its requirements were anti-innovation and negatively impacted the growth of India's AI ecosystem.[1926] This prompted India's IT minister to again post a message on X/Twitter, this time to clarify the scope of MEITY's advisory.[1927] Ultimately, the Indian government issued a revised two-page advisory on March 15, 2024, in supersession of the previous advisory. The new advisory removed the requirement of seeking government approval of generative AI models and submitting a compliance report.[1928]

This revised advisory, with an uncertain binding legal impact, was reportedly issued only to "significant" social media platforms (i.e., platforms with more than 500,000 registered users). It included various requirements, including that online platforms:

- not display or allow users to share unlawful content,
- not permit any bias or discrimination,
- not threaten the integrity of the electoral process,
- label AI-generated content that may include disinformation or deepfakes, and
- use metadata to identify any user that modifies information.

Additionally, under-tested or unreliable AI models could be released to Indian users only after implementing labeling mechanisms to flag the possible inherent fallibility or unreliability of the output generated. Users were to be informed through popups or other mechanisms about possible inaccuracies of AI-generated output.[1929] The advisory states that noncompliance with safe harbor protection laws for online platforms would result in legal consequences under existing Indian laws.

There is uncertainty regarding the legal basis and enforceability of such advisories, particularly given that MEITY lacks statutory authority to issue them. Additionally, the scope of the existing obligations that online platforms, including significant social media platforms, need to comply with to avail safe harbor protection under Indian law cannot be legally expanded through such advisories. Nevertheless, these advisories could be a possible indicator of what to expect in future regulations on AI that may be introduced by the Indian government.

5.4.3.B. Policy instruments promoting responsible and ethical AI

[1926] Suraksha P, Dia Rekhi & Annapurna Roy, *Govt missive to seek nod to deploy LLMs to hurt small companies: startups*, ECON. TIMES (March 4, 2024), https://economictimes.indiatimes.com/tech/technology/govt-missive-to-seek-nod-to-deploy-llms-to-hurt-small-companies-startups/articleshow/108185275.cms.

[1927] Rajeev Chandrasekhar (@Rajeev_GoI), TWITTER (March 4, 2024), https://twitter.com/Rajeev_GoI/status/1764577260647092368.

[1928] Aditi Agrawal, *In revised AI advisory, IT ministry removes requirement for govt permission*, HINDUSTAN TIMES (March 15, 2024), https://www.hindustantimes.com/india-news/in-revised-ai-advisory-it-ministry-removes-requirement-for-government-permission-101710520296018.html.

[1929] While both the advisories have not been officially released publicly by the Indian government, their copies are easily available on X/Twitter and LinkedIn.

Indian government bodies have undertaken various policy initiatives and efforts —such as publishing reports and fostering collaborations— to promote responsible AI practices and principles. These initiatives can be broadly categorized into two groups: federal efforts and those by state and other regulatory bodies.

1) Initiatives by federal government bodies

The federal government bodies have embarked on various initiatives.

a) Global Partnership on AI Summit

India hosted the Global Partnership on AI Summit in New Delhi in December 2023[1930] and is the Lead Chair of the Summit for 2024.[1931] One of the outcomes of the 2023 Summit was the "Ministerial Declaration 2023," in which member countries committed to continue working on advancing safe, secure, and trustworthy AI in their respective jurisdictions.[1932]

b) Report by the Economic Advisory Council to the Prime Minister on regulating AI

In January 2024, the Economic Advisory Council to the Prime Minister (EAC-PM) released a report titled *A Complex Adaptive System Framework to Regulate AI*.[1933] The report argues that AI systems must be treated as "complex adaptive systems" because AI systems behave like a "self-organizing entity" where individual algorithms follow basic protocols and then adapt in response to dynamic external environments.[1934] The report takes the example of how a facial-recognition algorithm may analyze billions of parameter tweaks chosen by a separate search algorithm. If the search algorithm evolves in unexpected ways, then it could dramatically impact the functioning of the final facial-recognition system.

The EAC-PM report recommends a regulatory approach that considers the nature of AI systems as "complex adaptive systems" based on the following governing principles:[1935]

1. Creating guardrails through setting boundaries and thresholds within which an AI system operates (such as not exceeding its intended functions) and implementing strict partitions and separate protocols for distinct AI systems through containerization or virtualization techniques. This would ensure that the harmful effects of one AI system do not cascade into a larger

[1930] *India brought together all major initiatives for AI – UN Advisory Group on AI, UK AI Safety Summit – at one event at GPAI New Delhi Summit*, PRESS INFO. BUREAU (14 Dec. 2023), https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1986475.

[1931] *GPAI Ministerial Declaration 2023*, GLOB. P'SHIP ON AI, https://gpai.ai/2023-GPAI-Ministerial-Declaration.pdf.

[1932] *Id.*

[1933] Sanjeev Sanyal, Pranav Sharma & Chirag Dudani, *A Complex Adaptive System Framework to Regulate Artificial Intelligence* (Econ. Advisory Council to the Prime Minister, EAC-PM/WP/26/2024, Jan. 2024), https://eacpm.gov.in/wp-content/uploads/2024/01/EACPM_AI_WP-1.pdf.

[1934] *Id.*, at 20.

[1935] *Id.*, at 21-25.

systemic failure, similar to a firebreak in forests. A regulatory sandbox can also help in achieving this outcome.

2. Providing points of human intervention in AI systems through manual override options, multi-factor authentication where certain decisions are subject to the approval of multiple persons, and "hierarchical governance" through multi-tier human approval processes.
3. Promoting transparency through the use of open-source licenses for core AI algorithms, open standards to facilitate audits and cross-system comparisons, standardized documentations of AI systems through "AI factsheets," and mandatory disclosures for extreme outcomes.
4. Ensuring accountability through legal frameworks that clearly define liability in case of malfunctioning or unintended outcomes from AI systems; delineating responsibilities among developers, operators and end users of AI systems; embed traceability mechanisms and mandate standardized incident reporting protocols.
5. Establish a specialized AI regulator with a centralized database for AI algorithms.

c) Recent initiatives by the Ministry of Electronics and Information Technology (MEITY)

In October 2023, the Ministry of Electronics and Information Technology released a report titled "India AI 2023," which contains observations from seven different working groups examining the promotion of different aspects of the AI ecosystem, such as establishing centers of excellence, dataset management, skilling, and computing infrastructure.[1936] The working group on dataset management provided a detailed framework for the establishment and functioning of a "National Data Management Office" (NDMO), previously proposed in the Draft National Data Governance Framework.[1937] The NDMO would essentially be responsible for the management of government datasets, which could then potentially act as an important asset for AI applications in India. Among other things, the NDMO would create standards based on principles of privacy by design, promote the use of privacy-enhancing technologies and systems, and adapt global best practices and standards as feasible to India. It would also create standards and principles for the ethical and fair use of data based on global standards.

MEITY announced a call for applications in December 2023, seeking project proposals from academic institutions, research and development organizations, and startups for building tools and frameworks on various "Responsible AI" themes. The themes included machine unlearning, synthetic data generation, algorithm fairness tools, AI bias mitigating strategies, ethical AI frameworks, privacy enhancing strategies, explainable AI (XAI) frameworks, AI ethical certifications, AI governance testing frameworks, and algorithmic auditing tools.[1938] The last date to submit proposals was February 4, 2024. This project is a part of the "Responsible AI" pillar of MEITY's National Programme on AI.

[1936] *India AI 2023: First Edition by Expert Group*, MINISTRY OF ELEC. & INFO. TECH., GOVT. OF INDIA (Oct. 2023), https://www.meity.gov.in/writereaddata/files/IndiaAI-Expert-Group-Report-First-Edition.pdf

[1937] *National Data Governance Framework Policy (Draft)*, MINISTRY OF ELEC. & INFO. TECH., GOVT. OF INDIA (May 2022), https://www.meity.gov.in/writereaddata/files/National-Data-Governance-Framework-Policy.pdf

[1938] *Call for Expression of Interest on Responsible AI*, MYGOV, GOVT. OF INDIA, https://innovateindia.mygov.in/eoi-responsibleai/.

d) Initiatives by the National Institution for Transforming India (NITI Aayog)

In 2021, the Indian government's official think tank, the Commission of the National Institution for Transforming India (also known as NITI Aayog), released a two-part series on responsible AI. Part 1 identified principles for responsible design, development, and deployment of AI in India.[1939] Part 2 set out enforcement mechanisms for these principles.[1940.] And in November 2022, the NITI Aayog released a case study on how the AI principles identified in its earlier reports could be applied to the deployment of facial recognition technology in India.[1941]

2) Efforts by other regulatory bodies and state governments

The Telecom Regulatory Authority of India (TRAI), an independent statutory body regulating certain aspects of telecommunication services in India, recommended, in July 2023, to establish an independent statutory authority, called the Artificial Intelligence and Data Authority of India (AIDAI).[1942] TRAI also recommended the formation of "a global agency that will act as the primary international body for development, standardization and responsible use of AI." However, TRAI's recommendations are not legally binding on the Indian government.

In 2019, the Securities and Exchange Board of India, the country's capital markets regulator, made it mandatory for various categories of regulated entities to report to the regulator the use of AI and ML technologies in their products and services.[1943] India's apex government medical research body released voluntary guidelines in March 2023 on the ethical use of AI in biomedical research and healthcare.[1944]

[1939] *Responsible AI #AIFORALL Approach Document for India, Part 1- Principles for Responsible AI*, NITI AAYOG (Feb. 2021), https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf.

[1940] *Responsible AI #AIFORALL Approach Document for India, Part 1- Operationalizing Principles for Responsible AI*, NITI AAYOG (Aug. 2021), https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf.

[1941] *Responsible AI #AIFORALL, Adopting the Framework: A Use Case Approach on Facial Recognition Technology*, NITI AAYOG (Nov. 2022), https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf.

[1942] *Recommendations on Leveraging Artificial Intelligence and Big Data in Telecommunication Sector*, TELECOMM. REG. AUTH. OF INDIA (July 20, 2023), https://trai.gov.in/sites/default/files/Recommendation_20072023_0.pdf.

[1943] Securities and Exchange Board of India, Reporting for Artificial Intelligence (AI) and Machine Learning (ML) applications and systems offered and used by market intermediaries, SEBI/HO/MIRSD/DOS2/CIR/P/2019/10 (Issued on Jan. 4, 2019), https://www.sebi.gov.in/legal/circulars/jan-2019/reporting-for-artificial-intelligence-ai-and-machine-learning-ml-applications-and-systems-offered-and-used-by-market-intermediaries_41546.html; Securities and Exchange Board of India, Reporting for Artificial Intelligence (AI) and Machine Learning (ML) applications and systems offered and used by Mutual Funds, SEBI/HO/IMD/DF5/CIR/P/2019/63 (Issued on May 9, 2019), https://www.sebi.gov.in/legal/circulars/may-2019/reporting-for-artificial-intelligence-ai-and-machine-learning-ml-applications-and-systems-offered-and-used-by-mutual-funds_42932.html.

[1944] *Ethical Guidelines for application of Artificial Intelligence in Biomedical Research and Healthcare*, INDIAN COUNCIL OF MED. RSCH., March 2023, https://main.icmr.nic.in/sites/default/files/upload_documents/Ethical_Guidelines_AI_Healthcare_2023.pdf.

Some state governments in India have also taken policy initiatives to promote responsible use of AI. For example, an Indian state government released a "Safe and Ethical AI Policy" in 2020.[1945] Another Indian state government signed a letter of intent with UNESCO in August 2023 to implement UNESCO's Recommendation on the Ethics of AI.[1946]

5.4.3.C. The Project of Digital India Act

In March 2023, MEITY released an *outline* for the Digital India Act, a law that would replace the existing Information Technology Act from 2000.[1947] The Indian government has yet to publicly release the first draft of the Digital India Act, though reportedly it has been holding consultations in cities such as Mumbai and Bengaluru.[1948]

The proposed law is intended to address challenges with the internet today, such as reining in the monopoly power of big tech companies while encouraging innovation, competition, and diversity through startup growth. Other challenges include addressing user harms, such as revenge porn, catfishing, doxxing, cyberstalking, and phishing; and trying to reduce hate speech and misinformation on the internet. While the outline of the proposed Digital India Act seems to be expansive in its scope as it seeks to cover different harms resulting from different technologies, it explicitly lists AI as one of the technologies it aims to regulate.

While the outline of the Digital India Act highlighted many issues with digital technologies like AI, the Indian government has not provided details about specific *regulations* to address these issues. In the outline's section regarding online safety and trust,[1949] MEITY has highlighted the need for "definition and regulation of high-risk AI systems through legal, institutional quality testing framework to examine regulatory models, algorithmic accountability…examine AI based ad-targeting, content moderation, etc." It also highlighted certain digital user rights, including the "right to be forgotten, right to secured electronic means, right to redressal, right to digital inheritance, right against discrimination, rights against automated decision-making, etc." MEITY also suggested that the Digital India Act could potentially eliminate the immunity granted to platforms under Section 79 of the IT Act, 2000. Section 79, similar to Section 230 of the U.S. Communications Decency Act, shields online intermediaries, such as social media and e-commerce platforms, from legal liability for third-party content posted on their platforms.[1950]

[1945] *Tamil Nadu Safe & Ethical Artificial Intelligence Policy 2020*, INFO. TECH. DEP'T, GOVT. OF TAMIL NADU (Oct. 2020), https://cms.tn.gov.in/sites/default/files/documents/TN_Safe_Ethical_AI_policy_2020.pdf.

[1946] *Unlocking the Potential of AI: UNESCO and Telangana Government spearheading ethics of AI for a better future*, UNESCO (August 23, 2023), https://www.unesco.org/en/articles/unlocking-potential-ai-unesco-and-telangana-government-spearheading-ethics-ai-better-future.

[1947] *Proposed Digital India Act, 2023*, MINISTRY OF ELEC. & INFO. TECH., GOVT. OF INDIA (Mar. 2023), https://www.meity.gov.in/writereaddata/files/DIA_Presentation%2009.03.2023%20Final.pdf.

[1948] *Digital India Act unlikely to be in place before LS election: Chandrasekhar*, BUS. STANDARD (Dec. 06, 2023), https://www.business-standard.com/india-news/digital-india-act-unlikely-to-be-in-place-before-ls-election-chandrasekhar-123120600305_1.html.

[1949] *Supra* note 1947, at 19.

[1950] Information Technology Act, 2000, §79.

Moreover, India's IT minister recently stated in an interview that the Indian government is working on a new AI law to protect the interests of news publishers and content creators, in addition to addressing user harms.[1951] This new law could be an independent law or be incorporated in the proposed DI Act.

Conclusion

Although the Indian government has yet to introduce a binding regulatory framework for AI-related products or services, AI has increasingly become a policy priority for India over the past five years. There are clear indicators of internal deliberations on a possible future regulatory framework to address the harmful consequences of AI-related applications, both generally and in specific sectors, such as health and finance. Additionally, there are developments at the intersection of AI and copyright-related concerns. Further policy and regulatory advancements on AI in India are likely to occur over the next year.

5.4.4 Israel

Protecting the tech sector's outsized role has, in no small part, motivated Israel to take a "soft,"[1952] risk-based, sector-specific regulatory approach to AI governance.[1953] The country's economy depends largely on a robust technology sector (the high-tech sector constituted about 18% of Israel's GDP in 2022).[1954] Israel's generative AI industry, in particular, has rapidly advanced: One report puts the country's generative AI venture capital ecosystem as the world's third largest.[1955]

Currently, there are no specific laws or regulations in Israel that directly govern AI. Israel does not have an AI-specific regulatory authority. The Ministry of Innovation, Science, and Technology (MIST) serves as the executive agency for national AI strategies and collaborates closely with the Ministry of Justice (MOJ). In 2022, MIST and MOJ published a draft policy document on AI.[1956] After conducting public consultations, both ministries released a policy paper in December 2023 entitled "Responsible Innovation: Israel's Policy on Artificial Intelligence Regulation and Ethics."[1957] This document outlines the country's

---

[1951] Ashwin Manikandan, *Exclusive: New AI law to secure rights of news publishers*, ECON. TIMES (Nov. 28, 2023), https://economictimes.indiatimes.com/tech/technology/exclusive-new-ai-law-to-secure-rights-of-news-publishers-ashwini-vaishnaw/articleshow/109043916.cms?from=mdr

[1952] Ministry of Innovation, Science, and Technology, *Israel's Policy on Artificial Intelligence Regulation and Ethics*, GOVERNMENT OF ISRAEL (December 17, 2023), https://www.gov.il/en/pages/ai_2023#.

[1953] *Id.*

[1954] *2023 Annual Report: The State of High-Tech*, ISRAEL INNOVATION AUTHORITY, https://innovationisrael.org.il/en/report/high-techs-contribution-to-the-economy/.

[1955] Elihay Vidal, *Israel's Generative AI map rapidly growing* (December 12, 2023), https://www.calcalistech.com/ctechnews/article/sk5rqxlua.

[1956] *Draft Policy of the Minister of Innovation, Science and Technology*, MINISTRY OF INNOVATION, SCIENCE AND TECHNOLOGY AND CONSULTING AND LEGISLATION DEPARTMENT (ECONOMIC LAW) AT THE MINISTRY OF JUSTICE (November 17, 2022), https://www.gov.il/en/pages/most-news20221117.

[1957] *Israel's Policy on Artificial Intelligence Regulation and Ethics*, MINISTRY OF INNOVATION, SCIENCE AND TECHNOLOGY AND CONSULTING AND LEGISLATION DEPARTMENT (ECONOMIC LAW) AT THE MINISTRY OF JUSTICE (December 17, 2023), https://www.gov.il/en/pages/ai_2023; *Regulations and Ethics, Responsible Innovation:*

approach to AI governance and policy. "Responsible innovation" is a term that captures the country's desire to protect and foster its growing tech industry while staying committed to non-binding global principles.

The policy paper recommends that regulators formulate their policies based on OECD principles to ensure the reliability of AI technology. Israel has officially endorsed the OECD's AI Principles, a set of internationally agreed-upon guidelines for AI actors (*see below section 6.2.1*).[1958] This approach aims to enhance growth, sustainable development, innovation, social welfare, and responsibility. Additionally, it emphasizes the importance of respecting fundamental rights and public interests, ensuring equality, preventing bias, and maintaining transparency, clarity, reliability, resilience, security, and safety.

The paper also recommends that Israel establish national guidelines to mitigate potential private sector abuses, such as discrimination, lack of human oversight, insufficient explainability, inadequate disclosure, safety issues, accountability gaps, and privacy violations. To address these concerns, the paper advocates avoiding broad horizontal legislation and operating,[1959] instead, within sector-specific regulations. This is in consideration of the diverse applications of AI technology, the limited understanding of its implications, and the rapid pace of technological advancements. Additionally, the paper advises adopting a flexible regulatory approach, including developing regulations in phases, using regulatory experiments, or implementing sandboxes.

In support of its advocated sector-specific approach, the paper calls for the establishment of an AI Policy Coordination Center under the Ministry of Justice. This center would function as an inter-agency body to facilitate coordination among various relevant agencies.[1960] This center would also be responsible for advising regulators, facilitating dialogue and knowledge-sharing with academia and industry, and helping regulators identify AI applications and challenges within regulated sectors. The paper also encourages other agencies that focus on digital policy issues to collaborate with each other to strengthen the country's sector-specific policies.[1961]

Conclusion

While there are currently no specific obligations imposed on developers, deployers, or users of AI systems, these actors should prepare for potential regulatory changes that may introduce legally binding or voluntary standards.[1962] The AI policy paper recommends that sectoral regulators develop appropriate regulations in alignment with OECD principles. In particular, the policy paper advocates for adopting a risk-based

Israel's Policy on Artificial Intelligence Regulation and Ethics (December 2023), https://www.gov.il/BlobFolder/policy/ai_2023/en/Israels%20AI%20Policy%202023.pdf.

[1958] *OECD AI Principles Overview*, OECD AI OBSERVATORY (May, 2024), https://oecd.ai/en/ai-principles.

[1959] Vidal, see *supra* note 1955.

[1960] *Id.*

[1961] Those agencies include the Israel Innovation Authority, the Privacy Protection Authority, the Israeli National Cyber Directorate, and the Israel National Digital Agency.

[1962] Daniel Turgel et al., *AI Watch: Global regulatory tracker - Israel*, WHITE & CASE (May 13, 2024), https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-israel.

approach through risk assessments conducted by the relevant sector-specific regulators, in line with the OECD principles.[1963]

5.4.5. Japan

Japan does not currently have a comprehensive law regulating the development and/or use of AI. Instead, the Japanese government has established non-binding guidelines that are generally applicable to these activities and has been promoting voluntary efforts by AI stakeholders. In July 2021, the Ministry of Economy, Trade, and Industry (METI) published a report stating, "From the perspective of balancing respect for AI principles and promotion of innovation, and at least at this moment, except for some specific areas, AI governance should be designed mainly with soft laws, which [are] favorable to companies that respect AI principles…." The paper added that "legally-binding horizontal requirements for AI systems are deemed unnecessary at the moment."[1964] This approach, based on non-binding guidelines, stems from the belief that binding laws cannot keep up with the rapid pace and complexity of AI development and might even stifle AI innovation. In February 2024, however, the Japanese authorities began discussions on the development of a binding law that would impose obligations on developers of large-scale foundational models.[1965]

Although Japan has not yet enacted binding legislation specifically targeting the development and use of AI, certain existing laws are applicable to these activities. For instance, Japan has developed guidelines to clarify how existing laws, such as those governing data protection and copyright, apply to the development, provision, and use of generative AI, thereby facilitating compliance for AI companies. Additionally, sector-specific frameworks regulate the deployment of AI in areas such as automobiles and medical devices.

This section will begin by exploring the Japanese government's provisional decision to adopt a non-binding framework. It will then analyze the application of existing personal data protection and copyright regulations to generative AI. Lastly, it will discuss the forthcoming regulatory framework currently under consideration by Japanese authorities.

5.4.5.A. The choice in favor of non-binding guidelines

Japan established the *Social Principles of Human-Centric AI* in 2019 to promote appropriate and proactive social implementation of AI.[1966] The document lists seven principles that AI stakeholders should keep in mind:

[1963] *Regulations and Ethics, Responsible Innovation: Israel's Policy on Artificial Intelligence Regulation and Ethics*, MINISTRY OF INNOVATION, SCIENCE AND TECHNOLOGY (Dec. 2023), https://www.gov.il/BlobFolder/policy/ai_2023/en/Israels%20AI%20Policy%202023.pdf.

[1964] 我が国のAIガバナンスの在り方 ver1.1 [AI Governance in Japan Ver.1.1], AIの原則の実践の在り方に関する専門家会議 [Expert Group on How AI Principles Should Be Implemented] (July 9, 2021), https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/2021070901_report.html.

[1965] 責任あるAI利活用を推進 AIPTが法制度の在り方検討 [Promoting Responsible Use of AI; AIPT Examines Legal System Framework], 自由民主党 [Liberal Democratic Party] (Feb. 26, 2024), https://www.jimin.jp/news/information/207671.html.

[1966] 人間中心のＡＩ社会原則 [Social Principles of Human-Centric AI], 人間中心のＡＩ社会原則会議 [Human-Centric AI Society Principles Council] (Mar. 29, 2019), https://www.cas.go.jp/jp/seisaku/jinkouchinou/index.html.

1. Human-Centric
2. Education/Literacy
3. Privacy protection
4. Ensuring security
5. Fair competition
6. Fairness, accountability, and transparency.
7. Innovation

On April 19, 2024, the Ministry of Internal Affairs and Communications and the Ministry of Economy, Trade, and Industry published the *AI Guidelines for Business*.[1967] The *AI Guidelines for Business* builds on the *Social Principles of Human-Centric AI* and outlines actions to be taken by AI developers, providers, and users (though this is not legally binding on them). Specifically, the *AI Guidelines for Business* includes 10 guiding principles:

1. Human-Centric
2. Safety
3. Fairness
4. Privacy protection
5. Ensuring security
6. Transparency
7. Accountability
8. Education/Literacy
9. Ensuring fair competition
10. Innovation

The *AI Guidelines for Business* then provides actions to be taken by AI developers, providers, and users with respect to each of the guiding principles. For example, AI developers are expected to take reasonable measures to control the quality of the data, noting that there may be biases in the training data (based on the Fairness principle). AI developers are expected to provide information to relevant stakeholders about the AI systems, such as information on safety and policies on collecting data learned by AI models (based on the Transparency principle). And AI developers are expected to prepare documents on the AI system development processes, the algorithms used, and the like, to the extent possible, in a form that can be used by third parties to validate the documents (based on the Accountability principle).

5.4.5.B. Application of the Data Protection Law to generative AI

Regarding generative AI and Japan's data protection law —called the Act on the Protection of Personal Information (APPI)— it is noteworthy that the Japanese Personal Information Protection Commission (PPC) issued administrative guidance to OpenAI and its operating company, on June 1, 2023, (PPC Guidance).[1968] The next day, the PPC issued a document entitled "Alert Regarding the Use of Generative AI Services" (PPC Alert).[1969] The PPC Guidance showed what businesses should be aware of when

[1967] AI事業者ガイドライン（第1.0版）[AI Guidelines for Business (Version 1.0)], 総務省 [Ministry of Internal Affairs and Communications] & 経済産業省 [Ministry of Economy, Trade, and Industry] (Apr. 19, 2024), https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240419_report.html.
[1968] 生成AIサービスの利用に関する注意喚起等について [Alert Regarding the Use of Generative AI Services], 個人情報保護委員会 [Personal Information Protection Commission](June 2, 2023), https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/.
[1969] *Id.*

developing generative AI, and the PPC Alert showed what businesses should be aware of when *using* generative AI.

1) Obligations of generative AI Developers

The PPC Guidance requests that OpenAI, in training AI models, not obtain sensitive personal information without obtaining the data subject's consent. Specifically, this requests OpenAI to

- make necessary efforts to ensure that sensitive personal information is not included in the training data, and
- not use, for machine learning, sensitive personal information entered by ChatGPT users who have opted not to have their personal data used for training AI models.

This follows the Act on the Protection of Personal Information (APPI) rule that says sensitive personal information (such as medical and criminal records) cannot be obtained without the consent of the subject of the data (APPI Article 20.2). But it should be noted that sensitive personal information made public *by* the subject of the data or by government agencies, media organizations, etc., is not subject to this rule (APPI Article 20.2.7).

Second, the PPC Guidance requests OpenAI to notify the data subject or make public the purpose of the use of the personal information it has obtained in Japanese; This follows the APPI rule requiring notification or publication of the purpose of use (APPI Article 21.1).

2) Obligations of users of generative AI

First, the PPC Alert requests businesses using generative AI services to enter personal information only to the extent necessary to achieve the specified purpose of use. This follows the APPI rule requiring that personal information be processed within the boundaries of its specified purpose of use, as stipulated in Article 18.1.

Second, the PPC Alert requests businesses to ensure that, when personal data are entered into generative AI services, the generative AI service providers do not use the entered data for purposes other than outputting responses, such as training AI models. If the providers do use the data for such purposes, the businesses are requested to obtain the data subject's consent. If, on the other hand, such personal data are not used for training AI models but only for the purpose of outputting responses to their instructions, the businesses are not requested to obtain the data subject's consent.

The APPI requires businesses to obtain the data subject's consent when they "provide" personal data to a third party, in accordance with Article 27.1. While entering personal data into a generative AI service would appear to constitute "providing" personal data to the generative AI service provider, the PPC Alert indicates that, if the personal data are used only to output responses to the business's instructions, it does not constitute "providing" and is not subject to the above rule.

5.4.5.C. Application of copyright law: copyright infringement and copyrightability

Regarding the Japanese Copyright Act and generative AI, it is necessary to pay attention to the *General Understanding on AI and Copyright (Copyright Guideline)* published by the Agency for Cultural Affairs

on March 15, 2024.[1970] The *Copyright Guideline* provides interpretive guidance under the Copyright Act on issues such as copyright infringement in the development and use of generative AI and whether the AI-generated products are protected as copyrighted works.

1) Obligation of generative AI developers

In collecting and processing training data and using it to train AI models, AI developers may use copyrighted works. The Japanese Copyright Act requires developers to obtain the permission of the copyright holder to use a copyrighted work in a way that infringes the copyright holder's rights, such as the right of reproduction. However, if the use falls under one of the exceptions, such as reproduction for private use or quotations, the developer does not have to obtain the permission of the copyright holder. The Japanese Copyright Act does not have a comprehensive exception limiting rights, such as Fair Use.

A 2018 amendment to the Copyright Act introduced an exception permitting the use of copyrighted works for "data analysis," including training AI models, as outlined in Article 30-4.2. This exception generally allows the use of copyrighted works to train AI models without the permission of the copyright holder. However, this exception does not apply when the purpose is to enjoy oneself or to have others enjoy the thoughts or emotions expressed in the copyrighted work. In this regard, the *Copyright Guideline* states that if an AI developer reproduces a copyrighted work in order to train an AI model for the purpose of intentionally outputting creative expressions contained in the training data, this could precisely fall under the purpose of enjoyment. The exception does not apply either when the use of the copyrighted work would "unreasonably prejudice the interests" of the copyright holder. According to the *Copyright Guideline*, this situation arises if an AI developer reproduces a copyrighted database work that is sold in the market without providing compensation.

If the use of a copyrighted work for AI learning constitutes copyright infringement, the copyright holder may seek damages and/or an injunction against the AI developer under Article 709 of the Civil Code and Article 112 of the Copyright Act. Regarding claims for injunctive relief, the *Copyright Guideline* notes that although requests to dispose of a trained model are generally not allowed, such a request may be permitted if the trained model frequently generates output similar to the copyrighted work in the training data.

2) Obligations of generative AI users

AI users may infringe on the copyright of an existing copyrighted work when generating and using output. Under Japanese law, copyright infringement is established if the generated output and the existing copyrighted work meet the following two conditions: 1) Reliance: the output was created based on an existing copyrighted work; and 2) Similarity: the creative expression is the same or similar. The *Copyright Guideline* states that, if the training data contain an existing copyrighted work and the AI generates output that is similar to that work, it is normally inferred that there is reliance between the copyrighted work and the generated output.

The *Copyright Guideline* also states that, in cases where the generation and use of output constitutes copyright infringement, copyright holders may, in principle, claim damages and injunctions only against

[1970] 文化審議会著作権分科会法制度小委員会 [Copyright Subcommittee, Legal System Working Group, Council for Cultural Affairs], AIと著作権に関する考え方について [On the Approach to AI and Copyright], 文化庁 [Agency for Cultural Affairs] (Mar. 15, 2024), https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/hoseido/r05_07/pdf/94024201_01.pdf; General Understanding on AI and Copyright in Japan-Overview, JAPAN COPYRIGHT OFFICE, AGENCY FOR CULTURAL AFFAIRS (May 2024), https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf.

AI users. However, in exceptional cases, copyright holders may also be able to take legal action against the AI developers and/or AI service providers. Specifically, the *Copyright Guideline* states that, if AI developers/service providers are aware that the AI model is likely to generate output similar to existing copyrighted works, but have not taken any measures to prevent this, copyright holders may have a claim against them.

3) Whether AI-generated products can be protected as copyrighted works

Regarding the protection of AI-generated products as copyrighted works, the principle is that if the AI user has creative intent and makes a creative contribution to producing the specific AI-generated product, then that product is protected as a copyrighted work.[1971] The *Copyright Guideline* states that if an AI user merely provides instructions for an idea, it is difficult to assert that they have made a creative contribution. However, if the AI user gives detailed instructions that clearly indicate creative expression, it is more likely that the AI-generated product will be protected as a copyrighted work.

5.4.5.D. Toward the adoption of a legal framework governing large-scale foundation models

In February 2024, the Liberal Democratic Party's AI Evolution and Implementation Project Team initiated discussions on drafting legislation to regulate the development of large-scale foundation models, provisionally titled The Basic Law for the Promotion of Responsible AI.[1972] The bill outlines its aim as follows: "To promote the development of an open environment that enables the design, development and introduction of safe, secure and responsible AI and the human-centered use of AI. The law aims to maximize the benefits of the sound development of AI, including innovation by AI, while minimizing the risk of violations of fundamental human rights and other rights and interests of the public through the utilization of generative AI and other AI."

The outline of the regulations presented in the statute is as follows:

(i) The government designates developers of AI foundation models that meet certain size and purpose criteria,

(ii) Designated developers must establish a system to fulfill specific obligations, including third-party verification of vulnerabilities and public disclosure of AI capabilities and limitations.

(iii) The private sector, including industry associations, must establish and publicize standards,

---

[1971] 知的財産戦略本部 [Intellectual Property Strategy Headquarters], 知的財産推進計画2023：多様なプレイヤーが世の中の知的財産の利用価値を最大限に引き出す社会に向けて [Intellectual Property Promotion Plan 2023: Towards a Society Where Diverse Players Maximize the Utilization Value of Intellectual Property in the World], 首相官邸 [Prime Minister's Office of Japan] (June 9, 2023), https://www.kantei.go.jp/jp/singi/titeki2/kettei/chizaikeikaku_kouteihyo2023.pdf.

[1972] 責任あるAI利活用を推進 AIPTが法制度の在り方検討 [Promoting Responsible AI Utilization: AIPT Examines Legal System Framework], 自由民主党 [Liberal Democratic Party] (Feb. 26, 2024), https://www.jimin.jp/news/information/207671.html; 衆議院議員　塩崎彰久（あきひさ） [Akihisa Shiozaki], 自民党AIの進化と実装に関するプロジェクトチーム [Liberal Democratic Party's Project Team on AI Evolution and Implementation], NOTE (Feb. 2, 2023), https://note.com/akihisa_shiozaki/n/n4c126c27fd3d.https://note.com/akihisa_shiozaki/n/n4c126c27fd3d.

(iv) Designated developers must regularly report their compliance status to the government or third parties, such as an AI Safety Institute,

(v) The government monitors and reviews designated developers based on their reports and requires them to implement remedies when necessary, and

(vi) The government imposes surcharges or penalties on designated developers for breaches of obligations.

In May 2024, the AI Strategy Council under the Cabinet Office —a meeting body established by the Japanese government to lead the rulemaking process for AI— began discussions on developing a binding legal framework for AI.[1973] The published document states the following:[1974]

> (i) Firstly, to promote the use of AI, the basic approach should be to make the maximum use of soft law to prevent overregulation of AI. However, for AI that is used in high-risk ways or that may lead to human rights violations, crimes, and other issues, it is necessary to consider the development of binding laws ("hard law").
>
> (ii) For AI with significant potential impact and risk if misused (such as large-scale foundation models), binding laws should be developed to complement soft law, and developers should be subject to obligations. However, as it is not easy to set uniform regulations, it is important to adopt a co-regulation approach where the government and the private sector collaborate to decide on the details of the regulatory operation.
>
> (iii) On the other hand, for AI with low potential impact and risk if misused, it is considered appropriate to avoid binding regulations and instead address the issues through soft law, such as the *AI Guidelines for Business*.

The "hard law" described in (ii) above is expected to be a statute along the lines of the Basic Law for the Promotion of Responsible AI, discussed above.[1975] Although discussions have only just begun and the future framework remains uncertain, Japan is clearly moving toward regulating the development of large-scale foundation models through binding legislation.

Conclusion

To date, Japan has addressed the risks associated with AI through two primary approaches: the establishment of non-binding guidelines generally applicable to AI development and use, and the application of existing sector-specific laws. Notably, Japan has developed guidelines on data protection and copyright laws to facilitate compliance for AI stakeholders. For instance, Japanese copyright law, in principle, allows the use of copyrighted works for training AI models without the permission of the copyright holder. In February 2024, the Japanese government initiated discussions on creating binding legislation that would impose obligations on developers of large-scale foundation models.

[1973] AI戦略実行会議（第9回）[AI Strategy Implementation Meeting (9th Session)], 内閣府 [Cabinet Office] (May 22, 2024), https://www8.cao.go.jp/cstp/ai/ai_senryaku/9kai/9kai.html..

[1974] AI戦略チーム [AI Strategy Team], 「AI制度に関する考え方」について [On "Approach to AI Systems"], 内閣府 [CAO] (May 2024), https://www8.cao.go.jp/cstp/ai/ai_senryaku/9kai/shiryo2-1.pdf.

[1975] 責任あるAI利活用を推進 AIPTが法制度の在り方検討 [Promoting Responsible AI Utilization: AIPT Examines Legal System Framework], 自由民主党 [Liberal Democratic Party] (Feb. 26, 2024), https://www.jimin.jp/news/information/207671.html.

5.4.6. Kingdom of Saudi Arabia

Focusing on growth and investment, the Kingdom of Saudi Arabia has not yet adopted a comprehensive legal framework for AI governance. However, the country has established a specialized authority and adopted ethical principles to guide AI development. Additionally, Saudi Arabia has updated its copyright and personal data laws to address the challenges posed by AI.

5.4.6.A. The Saudi Data Artificial Intelligence Authority (SDAIA)

In 2016, Saudi Arabia initiated a 14-year-plan titled "Vision 2030," an ambitious proposal from the Saudi Crown Prince, Mohammed Bin Salman.[1976] While this strategy emphasized rigorous investment in artificial intelligence,[1977] it led to the creation of the kingdom's chief regulatory authority, the Saudi Data Artificial Intelligence Authority (SDAIA), in August 2019. The same decree established the Saudi Artificial Intelligence Center and the Saudi Data Management Office, both of which fall under the SDAIA's purview.[1978]The SDAIA is responsible for overseeing the country's emerging AI research ecosystem and implementing new laws and guidelines that include an AI agenda.

The SDAIA's "National Strategy for Data & AI," which was released in October 2020, expanded the country's aggressive AI agenda to include, among many goals, the aim of having over 300 data and AI startups and to train and host over 20,000 data and AI specialists. Even so, most of the news and novelty surrounding Saudi Arabia and artificial intelligence stems from the country's ambitious investment strategy, which includes a $40 billion fund to invest in AI technologies.[1979]

5.4.6.B. The AI Ethics Principles

In September 2023, the SDAIA released the "AI Ethics Principles,"[1980] a set of principles with ambiguous guidance for local companies. The document enumerates seven principles with corresponding conditions

---

[1976]Adam Satariano & Paul Mozur, *To the Future': Saudi Arabia Spends Big to Become an A.I. Superpower*, THE NEW YORK TIMES (Apr. 25, 2024), https://www.nytimes.com/2024/04/25/technology/to-the-future-saudi-arabia-spends-big-to-become-an-ai-superpower.html.

[1977]Kingdom of Saudi Arabia, *Vision 2030*, https://www.vision2030.gov.sa/en/vision-2030/overview/#:~:text=Vision%202030%20creates%20a%20thriving,and%20prosperous%20future%20for%20all.

[1978]"*Saudi Arabia: Royal Order establishes Saudi Authority for Data and Artificial Intelligence*, ONETRUST DATAGUIDANCE (September 10, 2019), https://www.dataguidance.com/news/saudi-arabia-royal-order-establishes-saudi-authority-data-and-artificial-intelligence.

[1979] Saudi Arabia announced in March 2024 plans to create a $40 billion fund to invest in artificial intelligence. Maureen Farrell & Rob Copeland, *Saudi Arabia Plans $40 Billion Push Into Artificial Intelligence*, THE NEW YORK TIMES (Mar. 19, 2024), https://www.nytimes.com/2024/03/19/business/saudi-arabia-investment-artificial-intelligence.html

[1980] Saudi Data and AI Authority, *AI Ethics Principles*, KINGDOM OF SAUDI ARABIA (Sept., 2023), https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf.

necessary for their sufficient implementation. They include: fairness, privacy and security, humanity, social and environmental benefits, reliability and safety, transparency and explainability, and accountability and responsibility.

What makes the SDAIA's guidance slightly different from other ethical guidelines is that the SDAIA's appears mandatory for "adopting entities," which are defined as "Any public entity, business, or individual that is required to comply with the present document."[1981] Some say the guidance is, nonetheless, non-binding,[1982] while others have asked for clarification on the Principles' legal force.[1983]

Regardless, the new AI Ethics Principles constitute the principal regulatory framework for AI in the Kingdom of Saudi Arabia. Similar to the AI Act, the Principles categorize the risks associated with the development and utilization of AI into four levels with different compliance requirements for each:[1984]

- Little or No Risk: Systems classified as posing little or no risk do not face restrictions, but the SDAIA recommends compliance with the AI Ethics Principles.
- Limited Risk: Systems classified as limited risk are required to comply with the Principles.
- High Risk: Systems classified as high risk are required to undergo both pre- and post-deployment conformity assessments, in addition to meeting ethical standards and relevant legal requirements. Such systems are noted for the significant risk they might pose to fundamental rights.
- Unacceptable Risk: Systems classified as posing unacceptable risks to individuals' safety, well-being, or rights are strictly prohibited. These include systems that socially profile or sexually exploit children, for instance.

The SDAIA is responsible for monitoring compliance with the AI Ethics Principles, but the Principles also require adopting entities to employ four distinct employees to ensure an organization's compliance with the Principles.[1985] The responsibilities for these employees are vague in some instances, but they include: a Head of Entity/Chief Data Officer, who will oversee all elements of an AI ethics practice within an organization; a Chief Compliance Officer/Compliance Officer, who will serve as the "strategic lead" and personally see through the AI ethics practice's day-to-day operations; a Responsible AI Officer, who will serve as the operational lead of an organization's responsible AI practice and will work, specifically, with an organization's data management team to ensure compliance; and an AI Systems Assessor, who will audit AI systems and conduct periodic reviews to ensure compliance with the Principles.

[1981] *Id.*, at 30.

[1982] Brian Meenagh et al., *Artificial Intelligence Law: Saudi Arabia* in Artificial Intelligence Law, LATHAM & WATKINS LLP (January 17, 2024), https://www.lw.com/en/people/admin/upload/SiteAttachments/Lexology-In-Depth-Artificial-Intelligence-Law-Saudi-Arabia.pdf.

[1983] Nick O'Connell, et al., *Saudi Data and Artificial Intelligence Authority Reveals AI Ethics Principles 2.0*, AL TAMIMI & CO. (Oct. 23, 2023), https://www.tamimi.com/news/ai-ethics-principles-version-2-0/#:~:text=The%20framework%20is%20based%20on,%26%20Explainability%20and%20Accountability%20%26%20Responsibility.

[1984] Meenagh et al., *Artificial Intelligence Law: Saudi Arabia*, *supra* note 1982.

[1985] Saudi Data and AI Authority, *AI Ethics Principles*, *supra* note 1980.

5.4.6.C Amendments to copyright and data protection laws

Developments have also occurred in other areas of Saudi national law. The Saudi Authority for Intellectual Property (SAIP), a government authority for the regulation and protection of intellectual property, released a draft of amendments in April 2023 to the intellectual property laws for public consultation. This draft sought to integrate existing intellectual property regulations and added a specific section on AI. The new section, titled "AI-related IP and emerging technologies, and supporting their motivation," is aimed at regulating intellectual property issues related to AI and emerging technologies.

In March 2023, an updated version of the Personal Data Protection Law (PDPL) was issued. The revisions include added protections to the processing of a Saudi citizen's personal data by entities both residing inside and outside the country. While specific to data processing, the updated legislation, which is expected to go into effect in September 2024, still has restrictions for the processing of personal data by an AI system.

Conclusion

Currently the Saudi Data Artificial Intelligence Authority (SDAIA) is responsible for overseeing the country's AI ambitions and generating new regulations. Moreover, Saudi Arabia's "AI Ethics Principles" attempt to govern its growing AI industry. However, the legal force of the guidelines is unclear. At its strongest, it is a legally enforceable document. At its weakest, it is a non-binding set of guidelines.

5.4.7. Singapore

Singapore does not currently have binding AI legislation.[1986] It has frameworks which emphasize the country's commitment to a "soft law" approach, with non-binding instruments to provide guidelines for the ethical and responsible use of AI.[1987]

Much of Singapore's AI guidance comes from a single regulatory agency, the Infocomm Media Development Authority (IMDA), a "statutory board" or autonomous government agency, that operates under Singapore's Ministry of Communications and Information (MCI). The IMDA was initially a merger between two separate government agencies: one focused on the media sector and the other on the information and communications sectors.[1988] Today, the consolidated IMDA plays a leading role in establishing the country's tech policy and calls itself the "architects of Singapore's digital future."[1989]

There are specific government entities that also play important roles in the country's AI policy. The Smart Nation and Digital Government offices issued a comprehensive National AI Strategy in 2019 in which the

[1986] *Singapore's approach to AI governance regulation*, IAPP, https://iapp.org/resources/article/global-ai-governance-singapore/#regulation (last visited June 20, 2024).
[1987] Paulger, *supra* note 1326.
[1988] Agencies, Ministry of Communications and Information, https://www.mci.gov.sg/who-we-are/agencies/#:~:text=The%20Infocomm%20Media%20Development%20Authority,Infocomm%20Development%20Authority%20(IDA) (last visited June 20, 2024).
[1989] Infocom Media Authority, https://www.imda.gov.sg/about-imda/who-we-are (last visited June 20, 2024).

country made clear its vision to stimulate Singapore's economy and lead the global AI sector.[1990] As part of that Strategy, the government also introduced a National AI Office, within the Ministry, to carry out the country's ambitions.

In 2019, the IMDA issued the *Model AI Governance Framework* to provide actionable guidance for private organizations developing and deploying AI systems.[1991] In 2024, it published guidance specific to generative AI, through a document entitled *Model Governance Framework for Generative AI*. In the meantime, Singapore complemented both frameworks with applicable tools. Most notably, it created "AI Verify," an open source software product to test any AI products' compliance with international AI principles.[1992] Overall, the country has taken a non-binding approach to AI governance while emphasizing "practical" guidance with tools to make such guidance implementable.

5.4.7.A. The National AI Strategy (2019)

The National AI Strategy of 2019 articulated Singapore's ambitions to become a global leader in artificial intelligence. The Strategy suggested ways the country could better support and integrate its budding national AI industry within its larger economy. The first version of the document listed five national AI projects that the country would take steps to complete by 2030, including the integration of AI within freight planning, chronic disease management, and municipal services.[1993]

Singapore launched its second National AI Strategy, or NAIS 2.0, on December 2, 2023.[1994] This second initiative is more ambitious in scope and outlines 15 action items across three different "systems" for the country to undertake over the next three to five years.[1995] The action items include such things as enabling its growing network of AI researchers, businesses, and communities to better transform the national economy.

Although Smart Nation and Digital Government Office's National AI Office oversee implementation of these projects,[1996] the National AI Strategy stresses the important roles other government partners may play in the process. For instance, the document features AI Singapore as a key partner.[1997] Established in 2017

[1990] *Singapore National AI Strategy* (Nov. 2019), https://file.go.gov.sg/nais2019.pdf.

[1991] *Singapore's Approach to AI Governance*, PDPC (Nov. 3, 2023), https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework.

[1992] *Fact Sheet*, IMDA (June 7, 2023), https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2023/06/7-jun---ai-annoucements---annex-a.pPDPC | Singapore's Approach to AI Governancedf.

[1993] *Singapore National AI strategy* (Nov. 2019), see *supra* note 1990.

[1994] *National Artificial Intelligence Strategy 2.0 to Uplift Singapore's Social and Economic Potential*, SMART NATION SINGAPORE (December 4, 2023), https://www.smartnation.gov.sg/media-hub/press-releases/04122023/.

[1995] *Id.*

[1996] *Singapore's national AI strategy-1*, SINGAPORE MANAGEMENT UNIVERSITY, https://cityperspectives.smu.edu.sg/article/singapores-national-ai-strategy-1 (last visited June 20, 2024).

[1997] *Singapore National AI strategy* (Nov. 2019), *supra* note 1990.

by the National Research Foundation, AI Singapore (AISG) is a national research institute that serves as a national hub for AI research.[1998]

5.4.7.B. The Model AI Governance Framework (2020)

Singapore's first edition of its *Model AI Governance Framework (Model Framework), re*leased in 2019, invited public consultation and feedback.[1999] A year later, on January 21, 2020, the IMDA, after incorporating public comments, released a second edition of the *Model Framework*.[2000] The framework provides detailed guidance to private sector organizations –regarding ethical and governance issues– while implementing private sector AI solutions. It focuses largely on internal governance structures, human involvement in AI decision-making, operations management, and stakeholder interaction.

The *Model Framework* emphasizes two core principles: (a) AI decision-making should be explainable, transparent, and fair; and (b) AI solutions should prioritize human interests, enhance human capabilities, and safeguard human rights. It then uses these core principles to offer guidance across four domains: internal governance, human participation in AI-enhanced decision-making, operations management, and stakeholder engagement and communication. In addition, the framework provides that AI-related risks be managed within the existing enterprise risk management structures. This encompasses evaluating datasets for accuracy and bias risks and establishing comprehensive monitoring and reporting mechanisms.

The *Model Framework* serves primarily as guidance. Abiding by the framework's guidelines may help an organization comply with other national rules, such as the Singaporean data privacy laws. However, adoption of the framework is not alone sufficient to prove full compliance with, for instance, data privacy laws and other pieces of national legislation. One additional element the framework emphasizes is practicality. The IMDA said as much during the document's release: "Practicality is the hallmark of the *Model Framework*."[2001]

The IMDA issued companion documents to the *Model Framework*. These include, for instance, a two-volume "Compendium of Use Cases,"[2002] which looks in detail at the different organizations that were able to align their AI governance practices with those of the *Model Framework*. The IMDA also issued a Self-Assessment Guide for Organisations, which offers a way for organizations to assess how well they have

---

[1998] *AI Singapore*, AI SINGAPORE, https://aisingapore.org/ (last visited June 20, 2024).

[1999] *Singapore Releases Asia's First Model AI Governance Framework*, INFOCOMM MEDIA DEVELOPMENT AUTHORITY (Jan. 23, 2019), https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/archived/imda/press-releases/2019/singapore-releases-asias-first-model-ai-governance-framework.

[2000] *Model AI Governance Framework, Second edition (2020)*, INFOCOMM MEDIA DEVELOPMENT AUTHORITY AND PERSONAL DATA PROTECTION COMMISSION SINGAPORE (Jan. 29, 2020), https://www.symphonyai.com/wp-content/uploads/2023/05/SGModelAIGovFramework2-compressed.pdf.

[2001] *Singapore's Governing Framework for Artificial Intelligence*, N.Y. TIMES, https://www.nytimes.com/paidpost/imda/singapores-governing-framework-for-artificial-intelligence.html (last visited June 20, 2024).

[2002] *Singapore's Approach to AI Governance*, PDPC, https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework (last visited June 20, 2024).

implemented the *Model Framework*.[2003] Additionally, it provides an extensive list of useful industry examples and practices to aid organizations in implementing the Model Framework.

5.4.7.C. The Model AI Governance Framework for Generative AI (2024)

The IMDA and the AI Verify Foundation released a "Proposed Model AI Governance Framework for Generative AI" on January 16, 2024. This new framework builds upon the 2020 Model AI Governance Framework and follows a discussion paper titled "Generative AI: Implications for Trust and Governance."[2004] Following the publication of the proposed framework and feedback from various stakeholders, the finalized *Model AI Governance Framework for Generative AI* (*GenAI Framework*) was released on May 30, 2024, by the IMDA and the AI Verify Foundation.[2005]

The *GenAI Framework* aims to guide organizations in developing or deploying generative AI. It offers several recommendations for AI developers and companies, emphasizing safety, accountability, transparency, and security. Additionally, there are policy recommendations that, if adopted, could shift Singapore's approach from a soft law regime to a more binding framework, similar to those in other major jurisdictions.

The *GenAI Framework* develops a guidance that covers nine key areas:

1. **Accountability**: The *GenAI Framework* specifies that there must be a clear allocation of responsibility from both the start of the development process and once an AI system has been deployed.[2006] For issues that arise before the deployment process, the framework recommends allocating responsibility based on the level of control each stakeholder has within the generative AI development chain. For issues that arise after, it recommends indemnity and insurance to cover end users experiencing unanticipated issues. The framework also recommends policymakers create safety protections for end users, citing the EU's Revised Product Liability Directive and the AI Liability Directive (see above *section 5.1.3)* as potential models.[2007]

2. **Data Protection**: To ensure data protection, the *GenAI Framework* recommends clearly articulating how existing personal data laws apply to generative AI. It encourages the adoption of emerging privacy-enhancing technologies, such as anonymization techniques. And while the framework does not set a concrete legal solution to copyright protection

---

[2003] *Companion to the Model AI Governance Framework - Self-Assessment Guide for Organisations* (ISAGO), WORLD ECONOMIC FORUM (Jan., 2020), https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf.

[2004] *Generative AI: Implications for Trust and Governance*, SINGAPORE INFO-COMMUNICATIONS MEDIA DEVELOPMENT AUTHORITY (IMDA) & AICADIUM (June 23, 2023), https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf.

[2005] *Model AI Governance Framework for Generative AI*, IMDA (May 30, 2024), https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf.

[2006] *Id.*

[2007] *Id.*, at 6-7.

issues, it does urge further dialogue between policymakers and various stakeholders on the issue. The framework suggests, too, that developers adopt general best practices for data quality control and encourages a global effort to expand a pool of well controlled and diverse datasets for training.[2008]

3. **Trusted Development and Deployment**: The framework recommends adopting industry-wide best practices, which include Reinforcement Learning from Human Feedback, Retrieval-Augmented Generation, and "few-shot learning" (a machine-learning approach where models are designed to learn and generalize from a very limited amount of training data).

   Additionally, the *GenAI Framework* takes inspiration from the food industry and advocates for transparency measures that resemble a "food label" approach. This would require organizations to provide essential information, such as the data used, training infrastructure, evaluation results, safety measures, potential risks and limitations, intended use, and user data protection measures. These standardized disclosures could be voluntarily adopted by the industry or facilitated by governments and third parties.

   The *GenAI Framework* also recommends a more comprehensive and systematic approach to safety evaluations that goes beyond benchmarking and red teaming and accounts for other less talked about considerations, such as back-end safety. Even so, any standardized approach to safety will have to include a combination of baseline and sector-specific requirements. This includes an initial set of standardized model safety evaluations that will be regularly updated as advancements in the field occur.[2009]

4. **Incident Reporting:** The *GenAI Framework* recommends a number of reporting mechanisms to ensure sufficient and appropriate incident reporting. This includes, in addition to standard incident reporting practices, ensuring preemptive vulnerability reporting. The framework encourages AI developers to adopt the "Common Vulnerabilities and Exposures Program" (CVE), a tool used by software developers to create a shared list of known vulnerabilities.[2010] The framework also recommends adopting the EU AI Act's incident reporting requirement for reporting serious incidents.[2011]

5. **Testing:** The *GenAI Framework* speaks about encouraging third party testing and assurance. It recommends building common benchmarks and methodologies and other important measures for encouraging standardized testing practices. At the same time, it encourages an accreditation mechanism and cultivation of a pool of vetted third party auditors who will test AI systems.[2012]

[2008] *Id.*, at 8-9.

[2009] *Id.*, at 10-12.

[2010] *CVE® Program Mission*, CVE, https://www.cve.org/ (last visited May 22, 2024).

[2011] *Model AI Governance Framework for Generative AI*, *supra* note 2005, at 13-14.

[2012] *Id.*, at 15.

6. **Security**: The *GenAI Framework* recommends adopting a "Security-by-Design" approach in each stage of the systems development life cycle to address traditional security threats. It also recommends developing new tools, including input filters, digital forensics, and building out MITRE's ATLAS ("Adversarial Threat Landscape for AI Systems") database, an online knowledge base of tactics used against AI systems, to address novel threats specific to generative AI.[2013]

7. **Content Provenance**: To address concerns about deepfakes and disinformation, the *GenAI Framework* recommends both digital watermarking and utilizing cryptographic provenance solutions. However, the framework also suggests that these tools may not be feasible for all content creation. They may have limited effectiveness due to poor consumer understanding of the issue and the ability of malicious actors to circumvent these requirements. It recommends working with social media platforms and browsers to support the embedding and display of watermarks, standardizing the type of AI edits that need to be labeled, and raising awareness of content provenance for consumers.[2014]

8. **Safety and Alignment R&D:** The *GenAI Framework* recommends greater global cooperation to help research and development in alignment and to keep pace with model capabilities by establishing additional AI safety research and development institutions, like Singapore's Digital Trust Centre.[2015]

9. **AI for Public Good:** Finally, the *GenAI Framework* suggests that generative AI can have beneficial long-term effects by democratizing access to technology, delivering public services, upskilling the workforce, and contributing to sustainable growth through carbon footprint tracking.[2016]

While the *GenAI Framework* is comprehensive, it remains non-binding for AI companies. However, it includes several recommendations for policymakers. The framework advises updating existing laws to address emerging risks associated with AI use. For instance, policymakers should clarify current data protection laws to apply to generative AI and address complex intellectual property issues, such as the use of copyrighted material in training datasets. Additionally, the *GenAI Framework* emphasizes the need for a comprehensive safety evaluation framework for AI models. It suggests that Singaporean authorities should ensure greater transparency for higher-risk models, especially those with national security or societal implications. The framework also stresses the importance of enhancing global cooperation in AI model safety and alignment research and development.

[2013] *Id.*, at 16; MITRE ATLAS, https://atlas.mitre.org/ (last visited May 22, 2024).

[2014] *Model AI Governance Framework for Generative AI*, *supra* note 2005, at 17-18.

[2015] *Id.* at 19; *Digital Trust Centre*, NANYANG TECH. UNIVERSITY SING., https://www.ntu.edu.sg/dtc (last visited May 22, 2024).

[2016] *Model AI Governance Framework for Generative AI*, *supra* note 2005, at 20-21.

Overall, the *Generative AI Framework*, by emphasizing the need to update existing regulations to address emerging risks from AI use, may prompt future changes in Singaporean law. This indicates a potential shift away from Singapore's current "soft law" approach. It remains to be seen whether the Singapore government will choose to adopt similar approaches to those of other leading nations by enacting specific laws concerning AI.

Conclusion

Currently, Singapore has not chosen to regulate generative AI through binding legislation. Instead, it has adopted non-binding frameworks. The 2024 *Model AI Governance Framework for Generative AI* outlines principles and best practices that are widely accepted globally. This detailed and comprehensive framework is likely to significantly influence the practices of AI companies. The 2020 *Model AI Governance Framework* continues to apply to all AI systems.

Meanwhile, Singapore actively participates in global AI discussions and collaborations, contributing to the development of international AI standards. This involves working with organizations such as the World Economic Forum, the OECD, and the U.S. NIST to shape global AI governance. Given this context, it is possible that the Singaporean government may eventually decide to adopt AI laws, similar to that of the EU or China.

5.4.8. South Korea

South Korea has no specific laws or policies currently that directly regulate artificial intelligence. Its approach to AI regulation can be broadly summarized as "adopt first, regulate later."[2017] And overall, South Korea is currently maintaining a soft law approach to regulating AI but with potential legislation on the horizon.

Two of the primary agencies largely responsible for issuing regulatory guidance on AI are the Korean Ministry of Science and Information and Communication Technology (MSIT) and the Personal Information Protection Commission (PIPC). While PIPC is concerned with protecting and enforcing the country's privacy laws in the face of current and future AI risks, MSIT is largely responsible for guiding the country's AI initiatives.[2018] It is MSIT, for instance, that is housing the government's AI Strategy High-Level Consultative Council,[2019] a government-civilian partnership to discuss AI governance. And it was the MSIT Minister who announced, at the 2024 UK-South Korea AI Safety Summit, the creation of an AI Safety Institute.[2020]

[2017] Hwan Kyoung Ko, *Analysis of AI regulatory frameworks in South Korea*, ASIA BUSINESS LAW JOURNAL (April 15, 2024), https://law.asia/ai-regulatory-frameworks-south-korea/#:~:text=Although%20the%20details%20are%20yet,industrial%20activation%20of%20AI%20technologies.

[2018] Baek Byung-yeul, *Korea to invest $527 mil. to integrate AI into all sectors of society*, THEKOREATIMES (April 4, 2024), https://www.koreatimes.co.kr/www/tech/2024/06/129_372092.html.

[2019] *Id.*

[2020] Lee Dong-in et al., *Korea to establish AI safety institute at ETRI*, PULSE (May 23, 2024), https://pulse.mk.co.kr/news/all/11022500.

Although South Korea is not regulating AI at this time, the government has taken several measures to mitigate the potential harms of AI technology. These include steps that are similar to those taken by other countries:

- a national strategy ("AI National Strategy"),
- a set of guidelines ("Human Centered AI Ethics Standards"),
- a digital bill of rights ("Digital Bill of Rights"), and
- amendments to protect privacy (Personal Information Protection Act).

Korea has also had a bill pending in its National Assembly since 2021. While the bill, An Act on the Promotion of AI Industry and Framework for Establishing Trustworthy AI, (the AI Act),[2021] still awaits final votes from the National Assembly, it could potentially provide a legislative foundation for the country's AI strategy.

But any regulatory efforts will be balanced against the country's ambitions to "Leap to G3 in AI,"[2022] a slogan for the government's staunch aspirations to move from tenth place on the IMD's Digital Competitiveness rankings to third place before the end of the decade.[2023] As of 2024, South Korea was ranked sixth on the IMD ordering.[2024]

5.4.8.A. South Korea's AI National Strategy (2019)

The South Korean government in December 2019 introduced to the National Assembly the "AI National Strategy."[2025] The strategy, which is to be managed and implemented by the MSIT, outlines three core goals for the country to accomplish by 2030:

- significantly increase its overall digital competitiveness,
- expand AI's utilization across all sectors and professions, and
- improve overall quality of life by addressing AI's potential risks.

The AI National Strategy does not specify particular regulations but advocates for the creation and dissemination of an AI code of ethics.[2026] It underscores persistent concerns about a potential gap between the current legal system and emerging technologies, which could delay technological innovation due to the

---

[2021] Act on Promotion of the AI Industry and Framework for Establishing Trustworthy AI (Feb.14, 2023) https://www.assembly.go.kr/portal/bbs/B0000051/view.do?nttId=2095056&menuNo=600101&sdate=&edate=&pageUnit=10&pageIndex=1; Kim & Chang, *South Korea: Legislation on Artificial Intelligence to Make Significant Progress* (Mar. 6, 2023), https://www.kimchang.com/en/insights/detail.kc?sch_section=4&idx=26935.

[2022] *Leap to Global Top 3 in AI—Semiconductor, Advanced Biotechnology & Quantum Technology*, OFFICE OF THE PRESIDENT (Apr. 26, 2024), https://eng.president.go.kr/briefing/yKPaTKzX.

[2023] *National Strategy for Artificial Intelligence*, GOVERNMENT OF THE REPUBLIC OF KOREA (Oct. 28, 2019), https://wp.oecd.ai/app/uploads/2021/12/Korea_National_Strategy_for_Artificial_Intelligence_2019.pdf.

[2024] *IMD World Digital Competitiveness Ranking 2023*, WORLD COMPETITIVENESS CENTER, https://www.imd.org/centers/wcc/world-competitiveness-center/rankings/world-digital-competitiveness-ranking/ (last visited June 30, 2024).

[2025] *National Strategy for Artificial Intelligence*, see *supra* note 2023.

[2026] *Id.*, at 47.

absence of fundamental principles for addressing AI development. The strategy stresses the need for the current regulatory framework to shift to a negative regulatory system, which would allow all innovative attempts to create new services and would accelerate the spread of innovation by revising laws related to regulatory sandbox cases. The government envisions a "comprehensive negative list regulation roadmap"[2027] for the AI field, adhering to the basic principle of "Approval first and regulate later" to keep pace with rapid AI-based innovations. One of the strategy's key agenda items is "Drastic Regulatory Innovation and Revision of Laws," which includes the sub-task of "establishing framework legislation and reorganizing the legal system of each sector."[2028] The goal is to prepare framework legislation that defines a national strategic direction, incorporates the fundamental values and principles of the AI era, and implements measures to prevent dysfunction.

5.4.8.B. "Human-Centered AI Ethics Standards" and "Strategy to Realize Trustworthy AI"

In accordance with the AI National Strategy, MSIT released the "Human-Centered Artificial Intelligence Ethics Standards"[2029] in December 2020. Drawing on ethical guidance from the OECD, the EU, and Japan, these guidelines provide general principles applicable to any domain, issue, or technology. They include *three Basic Principles* for human-AI relations to achieve "Human-Centered AI," and *ten Requirements* that operationalize the Basic Principles.

The three Basic Principles are human dignity, the common good of society, and the development of AI to improve humanity ("Reasonableness of Technology").[2030] Among the Requirements that are to give effect to these three principles are guaranteeing human rights protections, respecting diversity by minimizing bias and discrimination, ensuring proper data management, and encouraging efforts to secure accountability, safety, and transparency.[2031] These standards are voluntary and intended to serve as a reference for all members of society. Yet, in February 2022, MSIT introduced a self-checklist and an AI Ethics development guide to support implementation across the private and public sectors.

On May 13, 2021, MSIT released its "Strategy to Realize Trustworthy AI," a detailed plan for implementing "Trustworthy AI for Everyone."[2032] The strategy lays out three principal objectives: create an environment for trustworthy AI, lay the foundation for AI's safe use, and spread AI ethics across society.[2033] Similar to the Standards, MSIT's strategy recommends ten tasks for MSIT to take by 2025 to ensure each of these

---

[2027] *Id.*, at 25.
[2028] *Id.*, at 24.
[2029] *The National Guidelines for AI Ethics*, S.KOR. INFO. SOC'Y DEV. INST. (Dec. 23, 2020), https://ai.kisdi.re.kr/eng/main/contents.do?menuNo=500011.
[2030] Paulger, *supra* note 1326.
[2031] *Id.*
[2032] *MSIT announced "Strategy to realize trustworthy artificial intelligence,"* MINISTRY OF SCIENCES AND ICT (May 13, 2021), https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=509&searchOpt=ALL&searchTxt.
[2033] *Id.*

objectives are met. These include activities such as ensuring sufficient vetting for trustworthiness at every stage of an AI's development and using safe and reliable data.[2034]

MSIT has not been the only Korean government agency to publish AI guidelines. Both the Korea Communications Commission (KCC) and the National Human Rights Commission of Korea (NHRC) have released their own AI frameworks. In 2019, the KCC established basic principles for user protection in algorithmic AI-based media recommendation services (ensuring that personalized recommendations are based on a uniform set of principles). And on May 11, 2022, the NHRC announced the "Human Rights Guidelines on the Development and Utilisation of AI."[2035] These Guidelines address issues including transparency and accountability in AI use, the necessity and methods for evaluating AI's impact on human rights, and the appropriate classification of AI based on risk levels.

5.4.8.C. The Digital Bill of Rights (2023)

MSIT released the "Charter on the Values and Principles of the Digital Society of Mutual Prosperity" on September 25, 2023.[2036] Known best by its short title, the Digital Bill of Rights,[2037] it is different from the Human-Centered Artificial Intelligence Ethics Standard in that the Digital Bill of Rights focuses more broadly on the values and principles that ought to underline South Korea's continued digital progress.

The ethical development of AI technologies is but one expectation of the Digital Bill of Rights. In general, the Digital Bill of Rights outlines five fundamental principles for which protection is necessary in order to achieve a "Digital Society of Mutual Prosperity." Those five principles include:

- "the guarantee of freedom and rights within a digital environment,
- the guarantee of fair access to and equitable opportunities in the digital world,
- building a safe and trustworthy digital society,
- promoting digital innovation based on autonomy and creativity, and
- advancing the well being for all humankind."[2038]

The document does not directly focus on providing ethical guidance for "trustworthy AI." Its primary concern is firmly establishing citizens' universal rights within a hazardous digital world. Simultaneously, it enumerates the obligations that states, private actors, and civil society each ought to follow in ensuring these universal rights are respected.

[2034] *Id.*

[2035] Byoung-il Oh, *The risks of artificial intelligence and the response of Korean civil society*, ASS'N FOR PROGRESSIVE COMMC'N (Mar. 5, 2024), https://www.apc.org/en/blog/risks-artificial-intelligence-and-response-korean-civil-society#:~:text=On%2011%20May%202022%2C%20the,impact%20assessment%20tool%20in%202024.

[2036] *South Korea presents a new digital order to the world!,* MINISTRY OF SCIENCES AND ICT (Sept. 25, 2023), https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=878&searchOpt=ALL&searchTxt=.

[2037] *Id.*

[2038] *Id.*

Some of the principles in the Digital Bill of Rights directly touch on popular issues in AI. "Building a Safe and Trustworthy Digital Society" is the goal that digital technologies are ethically developed.[2039] In addition, "the guarantee of freedom and rights" under the first principle would include the development of AI that is free of discrimination and bias.[2040] Furthermore, MSIT has said that the document is seen as fundamental to the ministry and the government's broader efforts to revise laws in anticipation of emerging technologies like AI.[2041]

5.4.8.D. Amendments to the Personal Information Protection Act (2024)

The Korean government introduced amendments March 14, 2023, to the Personal Information Protection Act (PIPA),[2042] South Korea's comprehensive data protection law.[2043] A year later, the Amended PIPA went into effect with substantial changes to address new AI capabilities.[2044] Among the most significant of those changes was the stipulation of data subjects' rights vis-a-vis automated decision-making. Part of the amended law stipulates that, when an AI-automated decision significantly impacts the rights or obligations of a data subject, the affected individual has the right to refuse the decision.[2045] At the very least, a data subject who is the object of a decision made through a fully automated process has the right to demand explanations or reviews of such decisions.[2046] In addition, entities that control data and meet threshold revenue or active user requirements must have a Chief Privacy Officer to ensure privacy protection as well as insurance coverage for any potential PIPA violations.[2047]

The Personal Information Protection Commission (PIPC), Korea's national data protection authority, is primarily responsible for crafting and enforcing data protection policies.[2048] Its duties encompass handling privacy-related complaints, engaging in international cooperation, and promoting education and technology dissemination regarding data privacy. The PIPC plays a vital role in safeguarding individuals' personal information and ensuring adherence to data protection laws. In 2023, the PIPC put forth AI-specific guidance to address potential privacy concerns.[2049] This document outlines the relevant privacy principles

---

[2039] *South Korea: Digital Bill of Rights - key takeaways*, DATA GUIDANCE (Feb. 2024), https://www.dataguidance.com/opinion/south-korea-digital-bill-rights-key-takeaways.
[2040] *Id.*
[2041] *South Korea presents a new digital order to the world!*, *supra* note 2036.
[2042] Personal Information Protection Act [Enforcement Date Sept. 15, 2023], https://www.pipc.go.kr/eng/user/lgp/law/lawDetail.do.
[2043] *South Korea - Data Protection Overview*, DATA GUIDANCE (June 2024), https://www.dataguidance.com/notes/south-korea-data-protection-overview.
[2044] Doil Son et al., *Amendment to the Enforcement Decree of the Personal Information Protection Act Comes into Effect*, LEXOLOGY (Mar. 2024), https://www.lexology.com/library/detail.aspx?g=2862aaf0-a64f-45b8-89fc-e6bee7596c44#:~:text=The%20Amended%20Enforcement%20Decree%20introduces,automated%20decisions%20facilitated%20by%20artificial.
[2045] *Id.*
[2046] *Id.*
[2047] *Id.*
[2048] Personal Information Protection Commission (PIPC), https://www.pipc.go.kr/eng/user/itc/itc/greetings.do (last visited June 20, 2024).
[2049] Paulger, *supra* note 1326.

and offers guidance on data protection obligations and best practices for each stage of an AI system's life cycle, including development (encompassing planning, data collection, and training) and deployment.[2050]

The PIPC has also played a surprisingly newsworthy enforcement role. In March 2023, the PIPC launched an investigation into OpenAI's ChatGPT following reports of personal data leaks. On July 27, 2023, the PIPC announced the investigation's findings and imposed a fine of KRW 3.6 million (approximately US $2,700). PIPC's investigation concluded that OpenAI failed to report a data breach that resulted in the leak of personal information of 689 South Korea ChatGPT users.[2051] PIPC also found OpenAI did not offer a Korean privacy policy and consent procedure in Korean. OpenAI also failed to comply with other legal provisions, by neglecting specific precautions to destroy personal data and allowing minors to register for ChatGPT.

5.4.8.E. Current proposals for AI regulation

South Korea is actively considering AI-specific legislation with its own so-called AI Act. Known more formerly as the Act on Promotion of AI Industry and Framework for Establishing Trustworthy AI, the AI Act[2052] is a consolidation of all previously introduced AI laws since 2021. The AI Act seeks to promote the AI industry while also protecting users by fostering a secure ecosystem through stringent notice and certification requirements. As outlined in the AI National Strategy, Korea aims to prioritize the technological development of its AI industry. Although regulation is a priority, the Act will support the country's burgeoning tech sector while safeguarding citizens from harmful risks.

Altogether, the proposal outlines principles for AI development and use, operator responsibilities, and user rights.[2053] It distinguishes prohibited, high-risk, and low-risk AI, aiming to establish a structured foundation for safe and trustworthy AI technology and policies.[2054] Public reporting about the AI Act indicates that it is guided by a few operating principles.[2055] First, the AI Act ensures that government pre-approval is not necessary for developing new technologies and generally supports business innovation in the AI industry. It sets forth categories of "high risk AI" that require an elevated level of trustworthiness. The AI Act also creates a statutory basis for ethical guidelines around AI.

The bill is currently the subject of intense debate, with the National Human Rights Commission of Korea (NHRC) and various human rights civic organizations submitting opinion letters. Additionally, a proposed

---

[2050] *South Korea: PIPC publishes guidance for the safe use of personal information in the age of AI*, DATA GUIDANCE (Aug. 3, 2023), https://www.dataguidance.com/news/south-korea-pipc-publishes-guidance-safe-use-personal.

[2051] Paulger, *supra* note 1326.

[2052] Act on Promotion of the AI Industry and Framework for Establishing Trustworthy AI (Feb. 14, 2023), https://www.assembly.go.kr/portal/bbs/B0000051/view.do?nttId=2095056&menuNo=600101&sdate=&edate=&pageUnit=10&pageIndex=1; Kim & Chang, *South Korea: Legislation on Artificial Intelligence to Make Significant Progress* (Mar. 6, 2023), https://www.kimchang.com/en/insights/detail.kc?sch_section=4&idx=26935.

[2053] Hwan Kyoung Ko, *Analysis of AI regulatory frameworks in South Korea* (Apr. 15, 2024), https://law.asia/ai-regulatory-frameworks-south-korea/.

[2054] Kim & Chang, *South Korea: Legislation on Artificial Intelligence to Make Significant Progress* (Mar. 6, 2023), https://www.kimchang.com/en/insights/detail.kc?sch_section=4&idx=26935.

[2055] *Id.*

amendment to Korea's Copyright Act, which aims to include explicit provisions for the use of copyright materials for data mining purposes, is also pending before the National Assembly. The Korean AI Act is currently under review by the National Assembly.

Conclusion

South Korea has announced ambitious goals to become a global leader in AI technology. In pursuit of that, its regulatory approach has been animated by the principle "permit first, regulate later." The National Assembly is discussing comprehensive AI legislation along with the country's own so-called AI Act. In the meantime, however, the MIST is seeing through the implementation of its own AI ethical standards and taking steps to carry out its "Strategy to Realize Trustworthy AI." Korean agencies, most notably the PIPC, have taken proactive measures to establish AI guidelines and enforce data protection measures on key AI players.

5.4.9. The United Arab Emirates

Among Arab countries in the Middle East, the United Arab Emirates (UAE) has taken a leading role in the development of artificial intelligence technologies. It was the UAE, with funding from the Advanced Technology Research Council under the Abu Dhabi government, that released the popular open-source large language model Falcon 180B (named after the UAE's national bird).[2056] When the UAE's Technology Innovation Institute publicly released Falcon in September 2023, Hugging Face hailed its arrival as the "largest openly available language model, with 180 billion parameters."[2057] Falcon's inception marked an important milestone in the UAE's national strategy to be a global AI leader.

Currently, there are no dedicated laws and regulations governing AI in the UAE. In October 2017, the country made clear its ambition to build its AI ecosystem under its "National Strategy for Artificial Intelligence 2031."[2058] The thrust of that strategy focuses on efforts to improve the UAE's AI competitiveness in both the region and the world. A less urgent priority under the document, albeit a priority nonetheless, is to "ensure strong governance and effective regulation." The "UAE Artificial Intelligence and Blockchain Council"[2059] is tasked with overseeing all aspects of the National Strategy, including the development of regulations and best practices on AI risks, data management, cybersecurity, and other digital issues.

[2056] Billy Perrigo, *The UAE is on a Mission to Become an AI Power*, TIME (Mar. 20, 2024), https://time.com/6958369/artificial-intelligence-united-arab-emirates/.

[2057] Philip Schmid et al., *Spread Your Wings: Falcom 180B is here*, HUGGING FACE (Sept. 6, 2023), https://huggingface.co/blog/falcon-180b.

[2058] UAE National Strategy for Artificial Intelligence 2031, UAE GOVERNMENT NATIONAL PROGRAM FOR ARTIFICIAL INTELLIGENCE (2018), https://ai.gov.ae/wp-content/uploads/2021/07/UAE-National-Strategy-for-Artificial-Intelligence-2031.pdf.

[2059] The UAE Artificial Intelligence and Blockchain Council is an entity established in 2018 to oversee and promote the integration and adoption of AI and blockchain technologies across various sectors in the United Arab Emirates, including finance, healthcare, transportation, and government services. This council operates as a collaborative platform that brings together experts, industry leaders, and government officials. UAE COUNCIL FOR ARTIFICIAL INTELLIGENCE AND BLOCKCHAIN, https://ai.gov.ae/ai_council/.

In furtherance of that objective, the City of Dubai launched the *AI Principles and Guidelines for the Emirate of Dubai* in January 2019.[2060] Dubai's AI Principles and Guidelines aim to foster the safe, responsible, and ethical development of AI by providing clarity to developers, government entities, and society. They promote fairness, transparency, accountability, and explainability in AI development and oversight. These principles seek to maximize innovation while minimizing societal risks, capturing economic and social benefits from AI to advance sustainability goals and expand its role in Dubai's economy.[2061] Dubai also introduced an 'Ethical AI Toolkit' outlining principles for AI systems to ensure safety, fairness, transparency, accountability, and comprehensibility.[2062] These guidelines could form the basis for future industry-specific policies in the UAE and beyond.

In January 2024, UAE's third president established an Artificial Intelligence and Advanced Technology Council to further design and implement AI policies.[2063] The Council will undertake the development and execution of policies and strategies concerning research, infrastructure, and investments in artificial intelligence and advanced technology. UAE is, likewise, notable for having the first AI minister position.[2064] This minister is responsible for directly supervising the state's AI ecosystem.

The UAE is a champion of "regulatory sandboxes," a strategy that involves limited live testing of a technology in a controlled environment under a regulator's direct supervision. The Emirates Regulations Lab, commonly known as RegLab, is an innovative initiative launched in January 2019 as a collaborative effort between the UAE government and the Dubai Future Foundation.[2065] The primary objective of RegLab is to create a proactive and adaptive legislative environment that keeps pace with rapid technological advancements. The Emirates Regulations Lab grants licensing for emerging and future technologies as part of its mandate to support and regulate innovation in the UAE. It provides licenses that allow for the controlled testing and development of new technologies in a real-world setting, for example for self-driving cars. These licenses are approved by the UAE Cabinet and are an integral part of RegLab's regulatory experimentation process. The UAE has said that its RegLab may grant temporary licensing for the testing and vetting of new AI technologies.[2066]

[2060] Digital Dubai, *Artificial Intelligence & Principles*, UAE GOVERNMENT (2023), https://www.digitaldubai.ae/initiatives/ai-principles-ethics#:~:text=AI%20ETHICS%20GUIDELINES,%2C%20transparency%2C%20accountability%20and%20explainability

[2061] *Id.*

[2062] Digital Dubai, *Artificial Intelligence Principles & Ethics* (2019), https://www.digitaldubai.ae/initiatives/ai-principles-ethics.

[2063] Andrea Benito, *UAE President announces the establishment of the AI and Advanced Technology*, CIO (Jan.23, 2024), https://www.cio.com/article/1297349/uae-president-announces-the-establishment-of-the-ai-and-advanced-technology.html.

[2064] Nick Fouriezos, *UAE tech minister: AI will be 'the new lifeblood' for governments and the private sector,* ATLANTIC COUNCIL (April 22, 2024), https://www.atlanticcouncil.org/blogs/new-atlanticist/uae-tech-minister-ai-will-be-the-new-lifeblood-for-governments-and-the-private-sector/.

[2065] Regulations Lab, *About Reglab*, UAE GOVERNMENT INITIATIVE, https://reglab.gov.ae/ (last visited June 20, 2024).

[2066] National Program for Artificial Intelligence, MINISTER OF STATE FOR ARTIFICIAL INTELLIGENCE OF THE UAE GOVERNMENT, https://ai.gov.ae/wp-content/uploads/2020/02/AIGuide_EN_v1-online.pdf (last visited June 20, 2024).

Within this context, OpenAI's Sam Altman declared at the UAE's AI Minister at the UAE's annual World Governments Summit that he believes the UAE, given its emphasis on experimental regulation, is well positioned to lead discussions on global AI regulations.[2067] Altman even suggests the United Arab Emirates could become a "regulatory sandbox" for artificial intelligence on the global stage.

Conclusion

Prioritizing innovation and competitiveness, the Emirates have not adopted a binding legal framework for AI regulation. Instead, they favor "innovation-friendly" regulatory sandboxes to foster technology development within a supervised environment. It is not surprising that this developer-friendly approach receives support from AI companies like OpenAI.

5.4.10. United Kingdom

The United Kingdom's approach to AI regulation has, in general, been one of a "light touch" so as not to run the risk of disrupting its developing AI ecosystem. Over the past few years, the UK has tried to position itself front and center of the global AI conversation, announcing a 10-year plan in 2021 to become an AI global superpower.[2068] The plan emphasizes a robust investment in research and development along with a governance framework that will prioritize innovation and risk management.

Former Prime Minister Rishi Sunak, in his October 2023 speech before The Royal Society, summarized the UK approach with a question: "How can we write laws that make sense for something that we don't yet fully understand?"[2069] He argued that he would not "rush to regulate" AI at this point in time.[2070] Nevertheless, the UK continues to issue guidance and safety precautions for AI developers and regulators. Reports in April 2024 indicated that the UK government could possibly be taking steps toward its own AI legislation despite the Prime Minister's earlier comments.[2071]

5.4.10.A. A pro-innovation approach to AI regulation

[2067] Abeer Abu Omar, *UAE Backs Sam Altman Idea to Turn Itself into AI Testing Ground*, BLOOMBERG (Feb.15, 2024), https://www.bloomberg.com/news/articles/2024-02-15/minister-backs-altman-s-idea-to-turn-uae-into-ai-testing-ground.

[2068] *National AI Strategy*, HM GOVERNMENT (Sept. 22, 2021), https://assets.publishing.service.gov.uk/media/614db4d1e90e077a2cbdf3c4/National_AI_Strategy_-_PDF_version.pdf.

[2069] Anna Gross, *Rishi Sunak says he will 'not rush to regulate' AI*, FINANCIAL TIMES (Oct. 26, 2023), https://www-ft-com.stanford.idm.oclc.org/content/509012f9-4e08-414c-a97f-dd733b9de6ef.

[2070] *UK PM Sunak warns against rush to regulate AI before understanding its risks*, ASSOCIATED PRESS (Oct.26, 2023), https://apnews.com/article/ai-artificial-intelligence-britain-sunak-7a7a90b4a94efd01e7a33bc3f75cd59b.

[2071] Anna Gross and Cristina Criddle, *UK rethinks AI legislation as alarm grows over potential risks*, FINANCIAL TIMES (Apr. 14, 2024), https://www-ft-com.stanford.idm.oclc.org/content/311b29a4-bbb3-435b-8e82-ae19f2740af9.

The UK's National AI Strategy, published on September 22, 2021,[2072] outlines a ten-year plan to position the UK as a global leader in artificial intelligence. The AI Action Plan, published on July 18, 2022,[2073] provides an overview of the activities being undertaken by various government departments to advance AI research, develop new AI applications, and strengthen the UK's AI capabilities.

Within this framework, the UK government has released various policy papers indicating that there are no immediate plans to introduce new AI-specific legislation. Instead, the government intends to direct existing regulators to interpret and implement the principles outlined in these policy documents.

1) UK's pro-innovation white paper (2023)

The UK government published its first AI Regulation Policy Paper on July 18, 2022.[2074] This document detailed the government's decision to adopt a regulatory framework that is "pro-innovation" and "context-specific." It outlined six cross-sectoral AI governance principles and indicated that the UK government intended to direct existing regulators to interpret and implement these principles across various sectors. The government then issued a white paper on March 29, 2023, entitled "A Pro-Innovation Approach to AI Regulation."[2075] Rather than creating new laws or establishing a standalone AI regulatory body, the white paper emphasized enhancing the authority of existing sector-specific regulators who already have the power to oversee AI within their respective domains.

Because the drafters of the white paper aimed to ensure relevance, adaptability, and flexibility, the document did not define artificial intelligence. Instead, the white paper concentrated on the two functions of AI technologies that most concern regulators: adaptivity and autonomy. "Adaptivity" addresses the worry that AI may evolve to a stage where the logic behind a generated outcome becomes difficult to discern. "Autonomy," on the other hand, pertains to the issue of assigning responsibility for decisions made by non-human entities. The government views this approach as a profound strength: "By defining AI with reference to these functional capabilities and designing our approach to address the challenges created by these characteristics, we future-proof our framework against unanticipated new technologies that are autonomous and adaptive."[2076]

The white paper tried to steer clear of overly restrictive rules for fear that premature and disproportionate regulation could dampen innovation. Instead, it focused on principles that could address needs as they arise.[2077] The white paper enumerated five non-binding principles for regulators to observe in regulating

---

[2072] *National AI Strategy*, *supra* note 2068.

[2073] *National AI Strategy - AI Action Plan*, HM GOVERNMENT (July 18, 2022), https://www.gov.uk/government/publications/national-ai-strategy-ai-action-plan.

[2074] *Establishing a pro-innovation approach to regulating AI*, HM GOVERNMENT (July 18, 2022), https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement.

[2075] *AI Regulation: A Pro-Innovation Approach*, HM GOVERNMENT (Aug. 3, 2023), https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#executive-summary.

[2076] *Id.*, see "3.2.1:Defining Artificial Intelligence."

[2077] *Id.*

emerging AI technologies. These principles are designed to ensure the responsible development and use of AI technologies while fostering innovation. The five principles are:

1. Safety, Security, and Robustness. Ensuring AI systems operate securely, reliably, and resiliently, minimizing risks to individuals and society.
2. Appropriate Transparency and Explainability: Promoting clarity and understanding of AI decision-making processes to enable accountability and trust.
3. Fairness: Addressing biases and ensuring that AI systems are fair and equitable, preventing discrimination and promoting inclusivity.
4. Accountability and Governance: Establishing clear responsibilities and governance structures for the deployment and use of AI systems.
5. Contestability and Redress: Providing mechanisms for individuals to challenge and seek redress for decisions made by AI systems.

These principles are variations of the OECD's AI Principles (*see below section 6.2.1*), a document to which the UK is a signed adherent. But where the OECD has a clear definition of an AI system, the UK does not. The UK instead favors a context-specific regulatory scheme. The white paper made clear the government's distinct approach: "We will not assign rules or risk levels to entire sectors or technologies. Instead, we will regulate based on the outcomes AI is likely to generate in particular applications."[2078] The government expects regulators to assess and apply the principles to AI use cases within their jurisdiction, prioritizing them according to sector-specific needs. Subsequently, regulators are anticipated to publish guidance on interpreting these principles within their respective domains, including providing practical tools to assist companies in complying with the principles. Finally, the government also announced its intention to establish a regulatory sandbox for AI, which, after an initial pilot phase, would expand to cover AI innovations across multiple sectors.

In November 2023, the UK government organized the inaugural international AI Safety Summit at Bletchley Park (*see below section 6.6)*. On this occasion, Prime Minister Rishi Sunak gave a pro-innovation speech in which he emphasized the values of ex post regulation and a "wait and see" approach.[2079]

2) Actions of UK regulators

In February 2024, the Department for Science, Innovation and Technology (DSIT) issued new guidance for regulators to assist them in interpreting and applying the principles-based approach.[2080] To ensure a coherent and streamlined AI regulatory landscape, the DSIT has established a Central AI Risk Function,

---

[2078] *Id*.

[2079] Ingrid Lunden, *At Bletchley, Rishi Sunak Confirms AI Safety Institute but Delays Regulations for Another Day*, TECHCRUNCH (Nov. 2, 2023), https://techcrunch.com/2023/11/02/at-bletchley-rishi-sunak-confirms-ai-safety-institute-but-delays-regulations-for-another-day/.

[2080] *Implementing the UK's AI Regulatory Principles, Initial Guidance for Regulators*, UK DEPARTMENT FOR SCIENCE, INNOVATION & TECHNOLOGY (Feb. 2024), https://assets.publishing.service.gov.uk/media/65c0b6bd63a23d0013c821a0/implementing_the_uk_ai_regulatory_principles_guidance_for_regulators.pdf.

which aims to enhance UK regulators' understanding of the AI risk landscape by providing expert risk analysis and supporting them in conducting risk assessments.

The government had requested that regulators update their regulatory approaches by May 2024.[2081] Several regulators have taken action. For instance, the Competition and Markets Authority (CMA) has published various reports and policy papers on AI and foundation models.[2082] Additionally, the Information Commissioner's Office (ICO), the UK's independent regulatory authority for data protection and privacy, released guidance in a brief April 2024 report.[2083] There, the ICO emphasizes how the government's principles align well with the ICO's data protection principles. The ICO also emphasizes that, in cases where personal data are processed, the ICO has the authority to intervene. The ICO requires AI organizations to mitigate risks to data protection and promote ICO's own risk mitigation tools. These include the AI and Data Protection Risk Toolkit,[2084] a set of best practices, and a Harms Taxonomy,[2085] which enumerates the different harms that an AI system may cause.

3) The UK government response to the consultation on AI regulation

Following up on the 2023 white paper, the UK government sought to further refine its regulatory framework for AI. It undertook a consultation process involving extensive stakeholder engagement, garnering over 545 responses from various sectors, including industry, academia, and civil society. On February 6, 2024, the UK government unveiled its response to the consultation.[2086] The response, led by the DSIT, reaffirmed the government's "pro-innovation" stance. The framework described in the government's response is principles-based, non-statutory, and cross-sectoral, aiming to balance innovation with safety by utilizing the existing regulatory framework for AI.

[2081] Jenna Rennie et al., *UK's Context-Based AI Regulation Framework: The Government's Response*, WHITE & CASE, https://www.whitecase.com/insight-our-thinking/uks-context-based-ai-regulation-framework-governments-response (last visited June 20, 2024).

[2082] *AI Foundation Models: Initial Review*, COMPETITION & MARKETS AUTHORITY (May 4, 2023), https://assets.publishing.service.gov.uk/media/64528e622f62220013a6a491/AI_Foundation_Models_-_Initial_review_.pdf; *AI Foundation Models: Initial Report*, COMPETITION & MARKETS AUTHORITY (Sept. 18, 2023), https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf; *AI Foundation Models: Update Paper*, COMPETITION & MARKETS AUTHORITY (Apr. 11, 2024), https://assets.publishing.service.gov.uk/media/661941a6c1d297c6ad1dfeed/Update_Paper__1_.pdf; *CMA AI Strategic Updates*, COMPETITION & MARKETS AUTHORITY (Apr. 29, 2024), https://www.gov.uk/government/publications/cma-ai-strategic-update/cma-ai-strategic-update#alt-text.

[2083] *Regulating AI: The ICO's strategic approach*, INFORMATION COMMISSIONER'S OFFICE (Apr. 30, 2024), https://ico.org.uk/media/about-the-ico/consultation-responses/4029424/regulating-ai-the-icos-strategic-approach.pdf.

[2084] *AI and data protection risk toolkit*, INFORMATION COMMISSIONER'S OFFICE, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/ (last visited on July 1, 2024).

[2085] *Overview of Data Protection: Harms and the ICO's Taxonomy,* INFORMATION COMMISSIONER'S OFFICE (April, 2022), https://ico.org.uk/media/about-the-ico/documents/4020144/overview-of-data-protection-harms-and-the-ico-taxonomy-v1-202204.pdf.

[2086] Jenna Rennie et al., *UK's Context-Based AI Regulation Framework: The Government's Response*, WHITE & CASE (Mar. 1, 2024), https://www.whitecase.com/insight-our-thinking/uks-context-based-ai-regulation-framework-governments-response.

The consultation response reaffirms the five initially proposed cross-sectoral principles. It also recognizes the need to expand these principles to explicitly address human rights, operational resilience, data quality, international alignment, systemic risks, broader societal impacts, sustainability, and education and literacy. These expanded principles are intended to guide the responsible design, development, and application of AI across various sectors. The response emphasizes a non-statutory, contextual approach, allowing existing regulators to apply the principles within their respective domains. However, the government acknowledges that most respondents to the consultation disagreed that implementing the principles through existing legal frameworks would effectively and fairly allocate legal responsibility for AI throughout its lifecycle.[2087] The government recognizes that, while immediate legislative action might be premature, future binding requirements may become necessary to address potential harms from advanced AI systems. The response points out the need for a deeper understanding of AI risks, regulatory gaps, and effective mitigation strategies, before pursuing legislative action.

While the government does not introduce or propose any new laws or regulations, it anticipates the need for legislation, particularly for General Purpose AI systems. To address this, it distinguishes between three types of AI technologies. These include:

- highly capable general-purpose AI, which encompasses foundation models that can perform a wide range of tasks and match or exceed the capabilities of today's most advanced models;
- highly capable narrow AI, which refers to foundation models designed to perform a limited or specific set of tasks within fields such as biology, with capabilities on par with or exceeding current advanced models; and
- agentic AI or AI agents, which are emerging technologies capable of competently completing tasks over extended periods and multiple steps.

Recognizing that highly capable general-purpose AI poses a significant challenge, the government plans to publish an update later this year to provide new responsibilities for developers of these systems.

In April 2024, early reports announced that policy officials from the DSIT were beginning to craft rules to regulate large language models.[2088] There is little information on what the scope and timing of the new regulations might be. Officials noted only that implementation of the new regulations would not be imminent and that it would apply only to large language models.[2089]

4) The AI Safety Institute

The UK government established the AI Foundation Model Taskforce in April 2023. Its primary goal was to create a team capable of evaluating the risks associated with advanced AI models. The task force was

---

[2087] *Id*.

[2088] Ellen Milligan, *UK Starts Drafting AI Regulations for Most Powerful Models*, BLOOMBERG (Apr. 15, 2024), https://www.bloomberg.com/news/articles/2024-04-15/uk-starts-drafting-ai-regulations-for-most-powerful-models.

[2089] Anna Gross & Cristina Criddle, *UK rethinks AI legislation as alarm grows over potential risks*, FINANCIAL TIMES (Apr. 14, 2024), https://www.ft.com/content/311b29a4-bbb3-435b-8e82-ae19f2740af9.

renamed the Frontier AI Taskforce in September 2023, when it published its initial progress report.[2090] Following the global AI summit in November 2023, the UK government announced that the task force would become a new AI Safety Institute (AISI).[2091] The institute is tasked with continuing the Frontier AI Taskforce's research and safety evaluations.

The AI Safety Institute (AISI) is a directorate of the DSIT.[2092] It aims to lead initiatives ensuring the safe development and deployment of artificial intelligence technologies. It particularly focuses on building internal capabilities to assess the safety of advanced AI systems, such as large language model assistants. The AI Safety Institute's goal is to conduct rigorous and reliable evaluations of these advanced AI systems both before and after their deployment.[2093] Moreover, the Institute must drive foundational AI safety research through exploratory projects. Additionally, it must collaborate with other international actors and national entities to provide up-to-date information on AI development and safety to the government.

5) The proposal of Artificial Intelligence (Regulation) Bill

While the government's approach to AI regulation has favored a "light touch," some policymakers in the UK have pushed for tougher regulations. Lord Christopher Holmes of Richmond, a member of the House of Lords Select Committee on Science and Technology, introduced The Artificial Intelligence (Regulation) Bill in November 2023, with the bill receiving a second reading in the House of Lords in March 2024.[2094] A key feature of the bill is the establishment of an AI authority that would coordinate across industries and agencies to ensure a unified, coherent, and aligned approach to AI regulation. The bill calls for the establishment of regulatory sandboxes for AI testing, as well as regular public consultations. While the bill has only a slim chance of passing into law,[2095] there are some who support specific features of the bill, such as the creation of greater regulatory alignment mechanisms.[2096]

5.4.10.B. The Generative AI Framework for His Majesty's Government (HMG)

In June 2023, in a report titled "Generative AI Framework for HMG," the UK's Central Digital and Data Office released specific guidance on generative AI for the UK civil service. The report enumerates 10

[2090] *First Progress Report*, FRONTIER AI TASKFORCE (Sept. 7, 2023), https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report.

[2091] *Introducing the AI Safety Institute*, DEPARTMENT FOR SCIENCE, INNOVATION, AND TECHNOLOGY (Nov. 2023), https://assets.publishing.service.gov.uk/media/65438d159e05fd0014be7bd9/introducing-ai-safety-institute-web-accessible.pdf.

[2092] AI Safety Institute, https://www.aisi.gov.uk/.

[2093] Emilia David, *UK mulling potential AI regulation*, THE VERGE (Apr. 15, 2024), https://www.theverge.com/2024/4/15/24131392/uk-ai-regulation-draft-safety.

[2094] Jedidiah Bracy, *Proposed UK AI regulation bill receives second reading in House of Lords*, IAPP (Mar. 25, 2024), https://iapp.org/news/a/proposed-uk-ai-regulation-bill-receives-second-reading-in-house-of-lords/#.

[2095] Lord Chris Holmes of Richmond MBE, *Artificial Intelligence (Regulation) Bill (AI Bill) Introduced*, LORDCHRISHOLMES (Nov. 30, 2023), https://lordchrisholmes.com/artificial-intelligence-regulation-bill/.

[2096] Bracy, *supra* note 2094.

principles to guide the responsible use of generative AI in government. They range from the strictly ethical ("use generative AI lawfully, ethically, and responsibly"[2097]) to practical recommendations ("understand how to manage the full generative AI lifecycle"[2098] and "have the skills and expertise needed to build and use generative AI"[2099]). The principles likewise emphasize the use of human control throughout the entirety of a generative AI product's life cycle, as well as protocols for keeping confidential and personally identifying data secure. Extensive as the document is, its preface states that the report is both "incomplete and dynamic"[2100] and emphasizes that guidance may change with new developments in generative AI.

While prioritizing the exercise of caution, the report also encourages the use of generative AI products to streamline government functions. In general, the report is largely written in plain, non-technical language with the first chapter serving as a primer on generative AI, a symbol of the government's intent to encourage civil servants to learn more.[2101] The government also lists strong and promising use cases for generative AI including: guiding digital inquiries, interpreting constituent requests, and improving search capabilities. Nonetheless, it stresses an abundance of caution when dealing with data, and it mandates the use of a Data Protection Impact Assessment ahead of deploying a generative AI technology that uses personal data. It also stresses the applicability of the UK General Data Protection Regulation, the Data Protection Act of 2018, and other data protection and privacy laws.

The "Generative AI Framework for HMG" notes the government's other efforts to support the responsible use of generative AI. It emphasizes existing guidelines that include the "Guidelines for AI procurement;" the "Digital, Data, and Technology (DDaT) Playbook;" the "Sourcing Playbook;" and the "Rose Book," which provides management guidance on knowledge assets. The report mentions existing regulations and policies that civil servants using generative AI need to be aware of, such as the UK Data Protection Act 2018, the Online Safety Act, and the "AI Assurance Techniques" from the Centre for Data Ethics and Innovation. The "Generative AI Framework for HMG" mentions the principles laid out in "A Pro-Innovation Approach to AI Regulation." In general, the report is meant to offer practical recommendations and case studies for civil servants to understand what the responsible use of generative AI looks like as a tool for carrying out important government functions.

5.4.10.C. The Online Safety Act 2023

The King, on October 26, 2023, gave his royal assent for the enactment of the Online Safety Act of 2023 (OSA), regarding content moderation.[2102] The law imposes a statutory duty of care on online services to mitigate harmful content for two categories of companies: 1) "regulated services" (user-to-user services which share user-generated content (e.g., Facebook) and 2) search services (e.g., Google). A company's obligations under the Online Safety Act will depend on the company's size and impact, with high-risk,

[2097] *Generative AI Framework for HMG*, CENTRAL DIGITAL AND DATA OFFICE (Jan. 18, 2024), https://www.gov.uk/government/publications/generative-ai-framework-for-hmg.

[2098] *Id.*, at 10.

[2099] *Id.*, at 12.

[2100] *Id.*, at 7.

[2101] *Id.*, at 13.

[2102] Online Safety Act 2023 (UK), c.50, https://www.legislation.gov.uk/ukpga/2023/50/enacted.

high-impact companies facing the strongest obligations.[2103] The OSA targets the mitigation of two categories of content: content that is illegal and content that is harmful to children.

The British Office of Communications (OfCom) enforces the Online Safety Act. Ofcom provides guidelines and a code of practice to help companies identify and manage illegal content. While companies are allowed to adopt their own approaches different from that of Ofcom, these alternatives must receive Ofcom approval. The fine for noncompliance with the OSA is £18 million (~$22 million) or 10% of a company's global revenue, whichever is higher.[2104]

The OSA specifies that any company that implements an approved code of practice or follows Ofcom's code of practice will not run the risk of penalty for illegal conduct even if illegal content is present on its services. In this way, the bill is more an incentive for clear and effective self-regulation than a constraint on companies.

Ahead of the law's passage, legislators added provisions that extended platform liability to include AI chatbot-generated content. The OSA has made it a criminal offense to share sexually explicit "deepfakes."[2105] Generative AI providers are not explicitly mentioned in the bill even though officials have made clear that the aim of the bill is to make technology companies responsible to their users.[2106] It will likely be left to the courts to determine how the statute will apply to generative AI more broadly. The law does not make AI companies responsible for illegal content produced or shared on their online service, but it does require them to establish internal policies for mitigating such content. To comply, companies must establish risk assessment processes to evaluate how often their services produce illegal content.

5.4.10.D. Intellectual property and generative AI

As in other jurisdictions, UK intellectual property law applicable to generative AI requires clarification, whether it regards the regime for data used to train models, the ownership of content produced with the help of generative AI, or the risks of infringing on a copyright.

1) Copyrights and training data

Similar to EU law (*see above section 5.1.1.B*), UK law includes statutory exceptions for Text and Data Mining (TDM) that permit the training of AI models using web-scraped data, though the scope of the UK law is narrower than Article 4(3) of the New Copyright Directive (EU) 2019/790. For example, "text and data analysis" is permitted for non-commercial research under section 29A of the Copyright, Designs and

---

[2103] *Online Safety Act: Everything We Know So Far*, AI, DATA & ANALYTICS NETWORK (Jan.4, 2024), https://www.aidataanalytics.network/data-governance/articles/online-safety-act-everything-we-know-so-far.

[2104] *Ofcom's approach to implementing the Online Safety Act*, OFCOM (Oct. 26, 2023), https://www.ofcom.org.uk/__data/assets/pdf_file/0017/270215/10-23-approach-os-implementation.pdf.

[2105] Alex Hern, *Online safety bill will criminalise 'downblousing' and 'deepfake' porn*, THE GUARDIAN (Nov. 24, 2022), https://www.theguardian.com/technology/2022/nov/24/online-safety-bill-to-return-to-parliament-next-month.

[2106] Jon Porter, *The UK just laid out new rules for the internet - it only gets harder from here*, THE VERGE (Nov. 8, 2023), https://www.theverge.com/2023/11/8/23952736/uk-online-safety-act-ofcom-illegal-harms-guidelines.

Patents Act 1988. In June 2022, the UK Intellectual Property Office (IPO) proposed expanding TDM exceptions to cover any purpose, including commercial uses, while maintaining its mandatory nature, thereby preventing rightsholders from opting out. However, due to significant opposition from influential stakeholders in the music and publishing industries, the UK Government announced in February 2023 that these proposals would be abandoned. Instead, it would develop, in consultation with a diverse group of experts, a code of practice. However, on February 6, 2024, in its response to the AI White Paper consultation (*see above section 5.4.10.A*), the UK government announced that it would ultimately not pursue the development of a code of practice concerning copyright and AI.

2) Copyrightability and patentability of AI generated outputs

UK Copyright law explicitly protects "computer-generated" works.[2107] The author of a computer-generated work is the person "by whom the arrangements necessary for the creation of the work are undertaken," as outlined by Section 9(3) of the Copyright, Designs and Patents Act 1988. Given this wording, it is unclear whether the user who inputs text *prompts* or the owner of the AI tool would be regarded as the author.

Moreover, the law also contains an "originality" criterion that requires literary, dramatic, musical, or artistic works (whether or not computer-generated) to be the "author's own intellectual creation"[2108] in order to be copyright protected. UK law has incorporated the "author's own intellectual creation" test introduced by the Court of Justice of the European Union (*see above section 5.1.1.B*). This requires that "the author was able to express their creative abilities in the production of the work by making free and creative choices so as to stamp the work created with their personal touch […] This criterion is not satisfied where the content of the work is dictated by technical considerations, rules or other constraints which leave no room for creative freedom."[2109] Where a literary, dramatic, musical or artistic work is created by a human (and involves human creativity) with minimal AI assistance (*e.g.*, music created with some AI assistance, but with a human choosing the instruments and tempo), the originality requirement should be satisfied.[2110] But computer-generated works that result from negligible human creative influence may not meet the originality requirement.

[2107] CDPA, § 9(3). "In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken." "Computer-generated" is defined as "work generated by computer in circumstances such that there is no human author of the work" (CDPA, § 178).

[2108] The European Court of Justice (CJEU) set it out in Infopaq International A/S v. Danske Dagblades Forening , case C-5/08, (July 16, 2009), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62008CJ0005. The originality test under English law was "skill, judgment and labor" until case law of the CJEU brought in the "author's own intellectual creation" test, which requires a higher standard of originality. In THJ Systems Ltd v. Sheridan [2023] EWCA Civ 1354, the UK Court of Appeal followed the CJEU test. However, there is no originality requirement for broadcasts, films, sound recordings, and published editions. As such, broadcasts, films, sound recordings, and published editions generated by AI would be protected without needing to consider the originality of these types of works.

[2109] Summarized by Arnold LJ in THJ Systems Ltd v. Sheridan [2023] EWCA Civ 1354, paragraph 16.

[2110] Assuming that the originality condition is met, the next question would be whether it is the AI system developer or user who undertook the "arrangements necessary for the creation of the work," and who, therefore, absent contractual provisions to the contrary, owns the copyright to the AI-created work.

As regards the patentability of AI-generated output, the UK Supreme Court considered, in Thaler v. Comptroller-General of Patents, Designs and Trademarks,[2111] whether an AI system could be an inventor under section 7(3) of the Patents Act 1977. The court held that the overall scheme of the Act assumes that patents can be granted to human persons only. And, on the facts of the *Thaler* case, the court decided that the owner of the AI system has no independent right to apply for a patent by virtue of simply owning the system. This decision does not, however, preclude the possibility of patent protection for an AI-generated output where a human is the inventor. In fact, the Supreme Court expressly stated (in *obiter* comments) that a patent application listing a natural person as an inventor who used an AI-system as a "highly sophisticated tool" may succeed.

3) Copyright infringement by AI-Generated outputs

Under section 16 of the UK Copyright, Designs and Patents Act 1988, copyright infringement occurs when a "whole or substantial part" of a particular work is copied. If an AI tool's outputs reproduce specific, identifiable sentences or images, this may constitute copyright infringement. However, this may be difficult to establish.

5.4.10.E. Other recent and potential developments

The UK's product-specific legislation, regarding such things as electrical and electronic equipment, medical devices, and toys, does not specifically address AI but may apply to some products that include integrated AI. In August 2023, the Department for Business and Trade and the Office for Product Safety and Standards published a consultation seeking views on proposals to overhaul the UK product safety framework. The consultation paper recognized that, as products become more sophisticated and driven by complex software, liability may not always be clear, particularly in relation to AI. So there may be developments in relation to product safety and AI in the mid- or long-term.

On the cybersecurity front, the UK government reported in "Safety and Security Risks of Generative Artificial Intelligence to 2025" that cyber-attacks are among the most significant risks that could manifest by 2025. On November 27, 2023, the UK National Cyber Security Centre published its "Guidelines for secure AI system development," which was developed with the United States' Cybersecurity and Infrastructure Security Agency.[2112]

Finally, under the National Security and Investment Act, 2021 (NSIA), certain investments (whether by UK or foreign investors) in businesses active in AI (and other sensitive sectors) require prior approval by the UK government. On April 18, 2024, the government published the outcome of its "Call for Evidence" on how the investment screening regime has been operating.[2113] A number of respondents welcomed clearer

[2111] Thaler v. Comptroller-General of Patents, Designs and Trademarks [2023] UKSC 49.

[2112] Guidelines for secure AI system development, NATIONAL CYBER SECURITY CENTRE (Nov. 27, 2023), https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development.

[2113] National Security and Investment Act 2021: Call for Evidence Response – Outcome, UK GOVERNMENT (Apr. 18, 2024), https://www.gov.uk/government/calls-for-evidence/call-for-evidence-national-security-and-investment-act/outcome/national-security-and-investment-act-2021-call-for-evidence-response.

definitions for the AI section of the NSIA. In response to this feedback, the government will launch a formal consultation on updating the definitions by summer 2024.

Conclusion

The UK government's strategy primarily focuses on promoting innovation. Instead of implementing blanket legislation, the UK government prioritizes a non-statutory, contextual, and cross-sectoral principles-based approach. For now, the UK has taken specific measures to support product safety, cybersecurity, and other areas. Additionally, it has called on federal agencies to submit guidance that aligns with the "Pro-Innovation Approach to AI Regulation" to comprehensively address significant AI risks. However, while the British government recognizes that binding requirements may eventually be necessary to mitigate potential AI-related harms, it has also stated that it will introduce legislation only when confident that such a step is warranted. Nevertheless, there are indications that the UK may be moving toward basic regulation.

KEY TAKEAWAYS

- The countries examined in this section can be classified into three distinct groups. First, several countries have clearly chosen to implement comprehensive legislation to regulate artificial intelligence, following the example of the European Union. Second, some countries, which initially dismissed legislative measures in favor of prioritizing innovation, are now revising their approaches and are seriously considering the adoption of a framework of binding legislation in the near to medium term. Finally, there are countries that reject any direct regulation of AI, opting instead to adopt non-binding ethical and technical guidelines.

- The countries currently engaged in an AI regulation process include Brazil and Canada. In Brazil, Bill 2338/2023, often referred to as the Brazilian AI Act, is expected to be adopted very soon. It appears to be influenced by the European Union's AI Act in several aspects, particularly by implementing a "risk-based approach" based on a gradation of risks. In Canada, Bill C-27, despite facing delays in its adoption, includes the Artificial Intelligence and Data Act, which draws inspiration from other frameworks, such as the EU AI Act and the U.S. NIST framework. Interestingly, in anticipation of adopting a binding legal framework, the Canadian government has published a non-binding code of conduct for firms that develop or manage generative AI with general-purpose capabilities.

- Some countries that initially did not choose to pass laws to govern AI are gradually moving towards adopting them. Japan serves as a particularly illuminating example. Initially, Japan opted for non-binding *guidelines* and explicitly excluded legally binding horizontal requirements for AI systems. However, the Japanese government is now considering the adoption of a binding legal framework, especially for high-risk AI systems and those with significant potential impact and risk if misused. Similarly, the Indian government initially issued "advisories" that imposed constraints on AI companies, such as the labeling of AI-generated content. However, the legal enforceability of these advisories remains uncertain. India is now contemplating the inclusion of provisions to regulate AI

systems, particularly high-risk ones, in the forthcoming Digital India Act, with the first draft expected to be released soon. South Korea initially embraced a soft law approach, following the principle of “permit first, regulate later.” However, potential legislation is on the horizon, with a future AI Act currently under review by the National Assembly. In Israel, where the focus has primarily been on promoting the tech sector, potential regulatory changes may introduce legally binding or voluntary standards. The government aims to avoid broad horizontal legislation and instead operate within sector-specific regulations.

- The third group of countries currently excludes the adoption of binding AI regulations. However, they still recognize the risks associated with AI and the importance of implementing guidelines to ensure adherence to key principles. The UK's approach is particularly noteworthy. Recognizing the risks and challenges posed by AI and the most advanced models, the UK government has championed several AI governance principles, taken a leading role in organizing international AI Safety Summits, and established an AI Safety Institute. Despite these efforts, the UK has so far excluded the adoption of a law to regulate AI, although it may reconsider this stance in the future. Singapore has clearly adopted a non-binding approach to AI governance, focusing on "practical" guidance with tools to facilitate implementation. Notably, its *Model AI Governance Framework* and *Model AI Governance Framework for Generative AI* offer actionable guidance for private organizations developing and deploying AI systems. Finally, Saudi Arabia has prioritized the adoption of AI Ethics Principles. Similarly, the United Arab Emirates does not anticipate adopting an AI law but champions “regulatory sandboxes,” which involve live testing of AI in a controlled environment under direct regulatory supervision.

Chapter 6 International initiatives and negotiations

Countries are not the only entities to establish standards and policies for the responsible governance of artificial intelligence. The profound implications of this rapidly developing technology make it a frequent subject of international discourse and negotiation by numerous, prestigious organizations and institutions. Experts emphasize that the potentially hazardous capabilities in the development and deployment of powerful, general-purpose AI systems generate significant global externalities.[2114] Consequently, international efforts to promote responsible AI practices are crucial for managing the associated risks. Various international organizations and multilateral institutions have initiated efforts to tackle the challenges and harness the opportunities presented by generative AI. For instance, the World Economic Forum, an international advocacy NGO and think tank, created the AI Governance Alliance to bring different stakeholders together to produce recommendations and regular reports.[2115]

However, experts are also still debating what international governance should look like and what entity, if any, should supervise compliance. What is undisputed is the need for developing and enforcing global rules and standards for AI —and specifically generative AI.. There have been various proposals for an AI-specific international institution. Some have argued for something equivalent to an International Atomic Energy Agency for AI,[2116] while others call for an Intergovernmental Panel on Climate Change-type organization (see above *section 1.3.2*).[2117] In a recent paper, several scholars have proposed a comprehensive set of governance functions at the international level to address AI-related challenges.[2118] These functions encompass supporting access to cutting-edge AI systems and establishing international safety standards.[2119]

This chapter does not delve into the detailed examination of the various policy proposals for establishing international AI governance, whether originating from experts, civil society, academics, or industry. Instead, it provides a general overview of the primary actions effectively undertaken by international organizations in this field, as well as the ongoing discussions at the international level. It is important to note that it does not cover *all* initiatives undertaken worldwide; only the most notable efforts.

[2114] Lewis Ho et al., *International Institutions for Advanced AI*, arXiv (July 11, 2023), http://arxiv.org/pdf/2307.04699.

[2115] World Economic Forum, AI GOVERNANCE ALLIANCE, https://initiatives.weforum.org/ai-governance-alliance/home (last visited June 29, 2024).

[2116] Jon Gambrell, *OpenAI CEO suggests international agency like UN's nuclear watchdog could oversee AI*, ASSOCIATED PRESS (June 6, 2023), https://apnews.com/article/open-ai-sam-altman-emirates-10b15d748212be7dc5d09eabd642ff39.

[2117] Suleyman & Schmidt, *supra* note 78.

[2118] Ho et al., see *supra* note 2114.

[2119] The paper discusses various institutional models, including a "Commission on Frontier AI," which would facilitate expert consensus on the opportunities and risks associated with advanced AI. Additionally, the scholars suggest a multi-stakeholder "Advanced AI Governance Organization" that would set international standards for managing global threats from advanced AI models, support their implementation, and potentially monitor compliance with a future governance regime. Another proposed model is the "Frontier AI Collaborative," aimed at promoting and disseminating access to advanced AI in underserved societies. Lastly, an "AI Safety Project" is envisioned to bring together leading researchers and engineers to study and mitigate technical AI risks. *Id.*

6.1. United Nations

The United Nations has been grappling with AI governance through its main bodies and many funds, programs, and specialized agencies. As AI holds the potential to impact the UN's core mandate of maintaining international peace and security, the organization has accelerated its efforts to both harness and control the technology. The International Telecommunication Union's 2022 report on the UN's AI activities detailed over 280 AI projects throughout UN entities.[2120]

AI governance work at the UN expanded significantly in 2023. The UN Security Council held its first meeting on the risks of AI in July 2023. At that meeting, Secretary-General António Guterres underscored the expectation that a multitude of governance responses from the international community will be necessary to address the complex economic and societal impacts of AI.[2121] In particular, Guterres emphasized the need to engage with problems from generative AI and to create flexible governance approaches that cover the technical, social, and legal questions surrounding the technology.[2122] Three flagship AI governance efforts stand out among the UN activities which align with this vision: the High-Level AI Advisory Body, the General Assembly AI resolutions, and the Global Digital Compact.

6.1.1. High-Level AI Advisory Body

The Secretary-General launched the High-Level Advisory Body on Artificial Intelligence in October 2023. The multi-stakeholder group is tasked with providing recommendations for the international governance of AI through two reports.[2123]

The Advisory Body released its first report, "Governing AI for Humanity," in December 2023. It detailed overarching recommendations for international governance,[2124] and emphasized that, while many past governance proposals have overlapping consensus, there is a lack of interoperability and alignment on implementation. The rapid development of AI technologies combined with a lack of coordination, threatens to enable new inequities in how the technology is both harnessed and regulated.

[2120] Int'l Telecommunication Union, *United Nations Activities on Artificial Intelligence (AI)*, ITU PUBLICATIONS (2023), https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2023-PDF-E.pdf.

[2121] Press Release, Secretary-General, *Secretary-General Urges Security Council to Ensure Transparency, Accountability, Oversight in First Debate on Artificial Intelligence* (July 18, 2023), https://press.un.org/en/2023/sgsm21880.doc.htm.

[2122] *Id.*

[2123] Press Release, Secretary-General, *UN Secretary-General launches AI Advisory Body on risks, opportunities, and international governance of artificial intelligence* (2023), https://www.un.org/sites/un2.un.org/files/231025_press-release-aiab.pdf.

[2124] Advisory Body on Artificial Intelligence, *Interim Report: Governing AI for Humanity*, UNITED NATIONS (Dec. 2023), https://www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf.

Rather than replicate an existing governance model for AI, the report articulated core functions for an AI governance entity that would be guided by respect for the UN Charter and international law. These include the need to:

- Assess and monitor AI directions, uses, and risks (Functions 1,6).
- Promote interoperability in governance frameworks and standards (Functions 2, 3).
- Facilitate international collaboration for the development and deployment of beneficial AI and its enablers (Functions 4, 5).
- Ensure compliance and accountability (Function 7).

The Advisory Body is set to release its second report in August 2024, which may provide further detail on a distinct form and timeline for a new international agency to govern AI based on these functions.[2125]

6.1.2. General Assembly AI resolutions

The UN General Assembly has adopted two resolutions on AI which emphasize international cooperation for safety and development. The General Assembly adopted the first in March 2024, focused on "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems" (Resolution A/78/L.49). [2126] The US led adoption of the resolution, of which 120 other member states co-sponsored. Though non-binding, the resolution emphasizes the role of international law in governing AI. It calls on member states and stakeholders to "refrain from or cease the use of artificial intelligence systems that are impossible to operate in compliance with international human rights law or that pose undue risks to the enjoyment of human rights."[2127]

In July 2024, the General Assembly adopted the second resolution for "Enhancing international cooperation on capacity-building of artificial intelligence".[2128] The Chinese-sponsored, non-binding resolution focuses on bridging gaps in AI development, calling on member states to institute capacity-building plans into their national AI strategies where possible. The resolution aims to foster a stronger environment of AI cooperation by encouraging more knowledge sharing, technology transfer, personnel training, and research collaboration within the international community.

6.1.3. Global Digital Compact

[2125] *Id.*

[2126] General Assembly resolution 78/49, *Seizing the opportunities of safe, secure, and trustworthy artificial intelligence systems for sustainable development*, G.A. Res. A/78/L.49 (Mar. 21, 2023), available at https://press.un.org/en/2024/ga12588.doc.htm

[2127] *Id.*

[2128] General Assembly resolution 78/311, *Enhancing International Cooperation on Capacity-building of Artificial Intelligence*, G.A. Res. A/RES/78/311 (July 5, 2024), available at https://www.un.org/en/ga/78/resolutions.shtml

The UN released the zero draft of its forthcoming Global Digital Compact (GDC) in April 2024.[2129] The Compact is set to be a governmental, but non-binding, guide to digital cooperation among UN-led multi-stakeholders..

One of the key objectives of the Compact is to improve international governance of emerging technologies, including AI. To enhance this international governance, the current draft (Revision 2)[2130] of the Compact makes multiple commitments for UN-led AI governance.[2131] These include commitments to:

- Establish an International Scientific Panel on AI.
- Initiate an Annual Global Dialogue on AI governance.
- Request the Secretary-General to establish a Global Fund for AI and Emerging Technologies.

The final form of the compact is set to be adopted at the September 2024 Summit of the Future.[2132] The outcomes of the Summit are, therefore, expected to set the stage for future UN-led AI governance initiatives.

Outside of these three main initiatives, other governance efforts could ramp up at the UN to impact AI. For example, the Secretary-General and the General Assembly's First Committee (on disarmament and international security) have warned that AI could enable risks from lethal autonomous weapons.[2133] Given the many agencies, funds, and programs throughout the UN system, specialized governance regimes could be developed which oversee specific sectors or applications of AI that have general impacts on international peace, security, and fundamental rights.

6.2. OECD

The Organisation for Economic Co-operation and Development (OECD) —a forum comprising 38 member countries— has assumed a leading role in global AI governance efforts. This international organization is dedicated to promoting policies that enhance the economic and social well-being of people worldwide. Its membership includes countries primarily from Europe, North America, and the Asia-Pacific region. The OECD engages in research and policy recommendations across various domains, such as economics, education, health, and the environment.[2134] Its work often involves developing standards and guidelines that shape global policies and practices. The forum also includes non-member "key partners," such as China,

[2129] United Nations, *Global Digital Compact: zero draft* (Apr. 1, 2024), https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/Global_Digital_Compact_Zero_Draft.pdf.
[2130] United Nations, *Global Digital Compact: rev. 2* (June 26, 2024), https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/GlobalDigitalCompact_rev2.pdf
[2131] United Nations, *Global Digital Compact: rev. 1* (May 15, 2024), https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/Global_Digital_Compact_Rev_1.pdf.
[2132] United Nations, *Global Digital Compact: Background Note* (Jan. 2023), https://www.un.org/techenvoy/global-digital-compact.
[2133] Press Release, General Assembly First Committee, *First Committee Approves Resolution on Lethal Autonomous Weapons, as Speaker Warns 'An Algorithm Must Not Be in Full Control of Decisions Involving Killing'* (Nov. 1, 2023), https://press.un.org/en/2023/gadis3731.doc.htm.
[2134]OECD, *Our History*, https://www.oecd.org/en/about/history.html (last visited July 15, 2024).

India, Indonesia, and South Africa. Collectively, OECD member countries and key partners control 80% of world trade and investment.[2135]

6.2.1. The OECD's recommendation on Artificial Intelligence

The OECD was an early mover in developing AI guidelines. It began to host AI-centric policy conferences in 2016,[2136] and two years later, its Committee on Digital Economic Policy (CDEP) gathered 50 global experts to draft ethical guidelines that would align artificial intelligence with human rights and democratic values.[2137] As a result of these consultations, the OECD adopted the official Recommendation on Artificial Intelligence,[2138] the world's first intergovernmental standard on AI, in May 2019.[2139]

The OECD Recommendation is one of the most cited AI guidelines.[2140] As the first of its kind, it serves as a foundational document for fostering innovation and building trust in AI. While non-binding, the document, nonetheless, carries political weight. All 38 OECD members and eight non-members, including Brazil, Egypt, and Singapore have signed on as adherents.[2141] The Recommendation served as the basis for the G20's AI Principles[2142] and has played a key role in legislative drafting of the European Union's AI Act and other national initiatives.[2143]

6.2.1.A. Content of the recommendation

The OECD's Recommendation has two primary sections. The first is a set of five foundational principles to ensure the "responsible stewardship of trustworthy AI."[2144] The second is a list of five policy recommendations for the responsible development of AI.

The five foundational principles are:

[2135] Hanni Rosenbaum & Ina Sandler, *Introducing Business at OECD,* INTERNATIONAL FEDERATION OF ACCOUNTANTS (Feb. 3, 2020), https://www.ifac.org/knowledge-gateway/discussion/introducing-business-oecd

[2136] Directorate for Science, Technology, and Innovation Committee on Digital Economic Policy, *Summary of CDEP Technology Insight Forum: Economic and Social Implications of Artificial Intelligence*, OECD Technology Foresight Forum 2016 on Artificial Intelligence (Nov. 17, 2016), DSTI/CDEP(2016)17, https://one.oecd.org/document/DSTI/CDEP(2016)17/en/pdf

[2137] OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (May 22, 2019) https://oecd.ai/en/assets/files/OECD-LEGAL-0449-en.pdf.

[2138] *Id.*

[2139] OECD, *AI Principles*, https://www.oecd.org/en/topics/ai-principles.html, (last visited July 15, 2024).

[2140] Nicholas Kluge Corrêa et al., *Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance*, 1 PATTERNS 2 (Oct. 13, 2023), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10591196/.

[2141] OECD, *AI Principles*, *supra* note 2139.

[2142] OECD, *What are the OECD Principles on AI?* (Mar. 2, 2020), https://www.oecd-ilibrary.org/what-are-the-oecd-principles-on-ai_6ff2a1c4-en.pdf.

[2143] Luca Bertuzzi, *OECD updates definition of Artificial Intelligence 'to inform EU's AI Act'*, EURACTIV (Nov. 14, 2023), https://www.euractiv.com/section/artificial-intelligence/news/oecd-updates-definition-of-artificial-intelligence-to-inform-eus-ai-act/.

[2144] OECD Recommendation, *supra* note 2137.

- **Inclusive growth, sustainable development, and well-being:** AI should advance the interests of people and the planet by fostering inclusive growth, sustainable development, and collective well-being.
- **Human-centered values and fairness:** AI systems must be designed to respect the rule of law, human rights, and democratic values. They should also include safeguards for fairness and justice and permit human intervention.
- **Transparency and explainability:** To help individuals recognize whether they are interacting with AI, AI systems should be sufficiently identifiable as such. They should also ensure transparency to allow human users to challenge negative outcomes.
- **Robustness, security, and safety:** AI systems must be reliable, safe, and secure throughout their life cycle, and organizations must continuously assess risks and manage the model throughout a given system's timeline.
- **Accountability:** Those developing, deploying, or operating AI systems should be held accountable for their actions in accordance with these principles.

Meanwhile, the five recommendations include:

- **Invest in AI research and development:** Encourage public and private investment in research and development for innovative and trustworthy AI.
- **Foster a digital ecosystem for AI:** Promote AI ecosystems that are accessible and include digital infrastructures, data, and knowledge-sharing.
- **Shape an enabling policy environment for AI:** Establish policy frameworks that support the deployment of trustworthy AI systems.
- **Build human capacity and prepare for labor market transformation:** Equip individuals with necessary AI skills and ensure a fair transition for workers.
- **International co-operation for trustworthy AI:** Enhance international and intersectoral cooperation to harmonize standards and approaches for trustworthy AI.

The Recommendation, in essence, sets out adaptive overarching principles to foster an environment conducive to trustworthy AI development. Recently, these principles have been updated.[2145] During the OECD's May 2-3, 2024, Ministerial Council Meeting, the organization included a number of revisions to its enumerated list of principles. Among the additions were explicit references to environmental sustainability and the importance of creating interoperability among the guidelines of different jurisdictions. The revisions also included increased emphasis on safety concerns and the need to address mis- and disinformation.[2146]

[2145] *OECD updates AI Principles to stay abreast of rapid technological developments*, OECD, (May 3, 2024), https://www.oecd.org/newsroom/oecd-updates-ai-principles-to-stay-abreast-of-rapid-technological-developments.htm.
[2146] *Id.*

6.2.1.B. The revised definition of AI

Four years after passing the first Recommendation, the OECD revised its definition of an "AI System," for adoption in the official EU AI Act.[2147] The Council of the OECD, the organization's overarching decision-making body, approved a revised definition of artificial intelligence on November 8, 2023 *(see above section 2.1.1)*.[2148] This updated definition describes an AI system as

"a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."[2149]

The new definition also provides that "different AI systems vary in their levels of autonomy and adaptiveness after deployment."[2150]

The revised definition reflects advancements in technology and market dynamics, aiming for international alignment, technical precision, and future readiness. Notably, it no longer requires AI objectives to be human-defined and acknowledges that systems can learn new objectives. With this new definition now official, it is being integrated into various pieces of legislation, such as the EU AI Act *(see above section 5.1.2.A (1))*.[2151]

As a token of the OECD's continued commitment to revising and clarifying its "AI system" definition, it published an "Explanatory Memorandum" in March 2024.[2152] The memorandum expands on each new word of the revised definition and notes that, despite the extensive work in defining an "AI system," there may be additional criteria to "narrow or otherwise tailor the definition when used in a specific context."[2153]

6.2.2. OECD's publications

The OECD publishes many reports focusing on generative AI that address various aspects of AI's impact, regulation, and governance. In April 2023, the OECD explored large language models in a policy brief that tried to explain natural language processing (NLP) and offered important policy considerations.[2154] The paper identified several pressing issues: the lack of explainability for internal NLP decisions, security, and safety risks; a potential increase in the proliferation of misinformation and disinformation; and, the lack of

---

[2147] Bertuzzi, *supra* note 2143.

[2148] Russell et al., *supra* note 84.

[2149] Grobelnik et al., *supra* note 86.

[2150] *Id.*

[2151] Bertuzzi *supra* note 2143.

[2152] OECD, *Explanatory memorandum on the updated OECD definition of an AI system*, OECD PUBLISHING (Mar. 5, 2024), https://www.oecd-ilibrary.org/science-and-technology/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.

[2153] *Id.*

[2154] OECD, *AI Language Models: Technological, socio-economic, and policy considerations*, OECD PUBLISHING (Apr. 13, 2023), https://www.oecd.org/publications/ai-language-models-13d38f92-en.htm.

minority language representation. The OECD also published another policy brief in September 2023, outlining the key issues that generative AI, as a whole, raises for policymakers. Among those issues: the amplification of misinformation and disinformation, the disruption of the labor market, and the interference with intellectual property rights.[2155]

Recent key publications about generative AI also include:

- "G7 Hiroshima Process on Generative Artificial Intelligence,"[2156] which presents the outcomes of the G7 initiative as of May 2023 (*see below section 6.3*);
- "Generative AI for Anti-Corruption and Integrity in Government: Taking Stock of Promise, Perils, and Practice."[2157]
- "The Future of Artificial Intelligence," Chapter 2 of the OECD Digital Economy Outlook 2024 (May 2024),[2158] provides insights into the advancements in AI technology and its future implications, as part of a broader analysis of the digital economy.[2159]

6.2.3. The AI Policy Observatory

The OECD created the OECD AI Policy Observatory[2160] to act as a type of full service research site for policymakers and AI experts, while promoting the OECD guidelines.

The Policy Observatory has been helping the OECD monitor the ways in which signers of the OECD Recommendation have abided by its principles and policy recommendations. The Observatory's biennial assessment, titled, "The State of Implementation of the OECD AI Principles,"[2161] reported that, as of 2024, over 50 countries had implemented national AI strategies, many of which directly referenced the OECD's principles.[2162] Of the 46 adherents to the OECD's AI principles, 41 had an AI strategy in place and three were in the process of developing one.[2163] The OECD report noted that, among member and non-member

---

[2155] Philippe Lorenz et al., *Initial Policy Considerations for Generative Artificial Intelligence*, OECD PUBLISHING (Sept. 18, 2023), https://www.oecd.org/publications/initial-policy-considerations-for-generative-artificial-intelligence-fae2d1e6-en.htm.

[2156] OECD, *G7 Hiroshima Process on Generative Artificial Intelligence (AI): Towards a G7 Common Understanding on Generative AI*, OECD PUBLISHING (Sept. 7, 2023), https://doi.org/10.1787/bf3c0c60-en.

[2157] Gavin Ugale & Cameron Hall, *Generative AI for anti-corruption and integrity in government: Taking stock of promise, perils and practice*, OECD PUBLISHING (Mar. 2024), https://doi.org/10.1787/657a185a-en.

[2158] OECD, *OECD Digital Economy Outlook 2024 (Volume 1): Embracing the Technology Frontier*, OECD PUBLISHING (May 14, 2024), https://doi.org/10.1787/a1689dc5-en.

[2159] *Id.*

[2160] OECD.AI Pol'y Observatory, *Background*, OECD.AI, https://oecd.ai/en/about/background (last visited Feb. 23, 2024).

[2161] OECD, *Report on the Implementation of the OECD Recommendation on Artificial Intelligence*, Doc. C/MIN (Apr. 24, 2024), https://one.oecd.org/document/C/MIN(2024)17/en/pdf.

[2162] *Id.* at 3.

[2163] *Id.*

countries, there were, as of May 2024, over 1,020 action programs initiated across 70 jurisdictions, a testament to the increased global attention to AI oversight since 2019.[2164]

Furthermore, the OECD's AI Policy Observatory site features a growing live repository that tracks the AI regulatory landscapes of 69 different countries and territories.[2165] It also provides tools for auditing AI systems[2166] and a recently launched Global AI Incident Monitor (AIM).[2167] In further support of the AIM and the OECD's work gathering reports on AI incidents, the AI Policy Observatory published a report titled, "Defining AI incidents and related terms" that offers important definitions and distinctions between "AI incidents" and "AI hazards."[2168]

6.3. The G7

The Group of 7 has been at the center of global efforts to regulate AI. Comprised of the European Union and seven of the world's most economically advanced countries (Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States), the G7 gathered at the Hiroshima Summit in May 2023, in part to discuss an AI agenda.[2169] The summit launched the eponymous *Hiroshima AI Process*, a ministerial forum for G7 ministers to discuss AI governance and collaborate toward an international framework.

The specific components of such a framework remained undetermined. In the following month, the G7 distributed a survey to its members to solicit their perspectives on the appropriate direction. The OECD, a frequent G7 partner, gathered countries' responses to the survey and published a report on September 7, 2023, titled the *G7 Hiroshima Process on Generative Artificial Intelligence (AI)*.[2170] It identified the G7's most urgent priority as "responsible use of generative AI technologies"[2171] and offered, as one potential solution, the creation of a voluntary code of conduct.

The OECD report was released during the G7 Digital and Technology Ministers' discussions on the future of the *Hiroshima Process*. Following the meeting, the G7 Digital and Tech Ministers announced they

---

[2164] *Id.*

[2165] OECD.AI Pol'y Observatory, *National AI Policies & Strategies*, OECD.AI, https://oecd.ai/en/dashboards/overview (last visited Feb. 23, 2024).

[2166] OECD.AI Pol'y Observatory, *Catalogue of Tools & Metrics for Trustworthy AI*, OECD.AI, https://oecd.ai/en/catalogue/tools?terms=audit&approachIds=1&objectiveIds=2&orderBy=dateDesc (last visited Feb. 23, 2024).

[2167] OECD.AI Pol'y Observatory, *OECD AI Incidents Monitor (AIM)*, OECD.AI, https://oecd.ai/en/incidents (last visited Feb. 23, 2024).

[2168] OECD, *Defining AI incidents and related terms*, OECD PUBLISHING (May 6, 2024), https://doi.org/10.1787/d1a8d965-en.

[2169] G7 Hiroshima Leaders' Communiqué, (May 20, 2023) https://www.mofa.go.jp/policy/economy/summit/hiroshima23/documents/pdf/Leaders_Communique_01_en.pdf?v20231006.

[2170] OECD, *G7 Hiroshima Process on Generative Artificial Intelligence*, *supra* note 2156.

[2171] *Id.* at 15.

endorsed four initiatives in furtherance of the *Hiroshima Process*:[2172] 1) the then already released OECD report, 2) international guiding principles for AI actors, 3) a code of conduct for organizations developing advanced AI systems, and 4) several cooperative research projects that would support AI tools and best practices. Together, these elements would form the *Hiroshima AI Process Comprehensive Policy Framework*,[2173] a non-binding rulebook for the responsible development of generative AI.

With surprising speed, the G7 began work in May 2023 and five months later, on October 30, 2023, released a *Code of Conduct* and separate *Guiding Principles* for organizations developing advanced AI systems."[2174] On December 1, 2023, the G7 Digital and Tech Ministers held another meeting, this time to formally agree to the updated *Hiroshima AI Process Comprehensive Policy Framework*, the first AI framework of its kind.[2175] In this version, the first three elements of the *Framework* were kept with few amendments. The fourth element (cooperative research projects) was specified to include research projects studying content authentication and labeling mechanisms, among other tools.[2176]

6.3.1. Content of the *Hiroshima AI Process Comprehensive Policy Framework*

The *Hiroshima AI Process Comprehensive Policy Framework* consists of four elements:

- The OECD report, *G7 Hiroshima Process on Generative Artificial Intelligence* (AI),
- *The Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems*,
- *The Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*, and
- Joint research projects on generative AI with GPAI and the OECD.

The Guiding Principles and the *Code of Conduct* form the bulk of the *Framework* and are the two elements that receive the most attention. Both documents underscore the G7's strong commitment to key aspects of AI governance, addressing the "design, development, deployment and use of advanced AI systems."[2177] They incorporate a wide array of existing international principles, offering a more detailed set of guidelines for AI actors compared to the OECD AI Principles, to which they explicitly refer.[2178] This set of guidelines

---

[2172] G7 Hiroshima AI Process, *G7 Digital & Tech Ministers' Statement* (Sept. 7, 2023), https://www.politico.eu/wp-content/uploads/2023/09/07/3e39b82d-464d-403a-b6cb-dc0e1bdec642-230906_Ministerial-clean-Draft-Hiroshima-Ministers-Statement68.pdf.

[2173] JapanGov, *The Hiroshima AI Process: Leading the Global Challenge to Shape Inclusive Governance for Generative AI*, KIZUNA (Feb. 9, 2024), https://www.japan.go.jp/kizuna/2024/02/hiroshima_ai_process.html#:~:text=Amid%20the%20growing%20global%20debate,%2C%20secure%2C%20and%20trustworthy%20AI.

[2174] G7 Leaders' Statement on the Hiroshima AI Process, (Oct. 30, 2023) https://www.mofa.go.jp/ecm/ec/page5e_000076.html.

[2175] Hiroshima AI Process, https://www.soumu.go.jp/hiroshimaaiprocess/en/index.html (last visited June 20, 2024).

[2176] *G7 Digital & Tech Ministers' Statement, supra* note 2172.

[2177] Hiroshima AI Process, *International Code of Conduct for Organizations Developing Advanced AI Systems*, https://www.mofa.go.jp/files/100573473.pdf (last visited June 20, 2024).

[2178] *Id.*

includes a risk-based approach applied throughout the AI lifecycle, starting with precautionary pre-deployment risk assessments and mitigation strategies. AI developers and deployers are required to implement risk management policies and procedures, along with robust security controls, including internal adversarial "red teaming" exercises. The *Guiding Principles* and the *Code of Conduct* also stress the importance of continuous monitoring, reporting, and mitigation of misuse and incidents. Additionally, they identify priority areas in AI research and development, such as content authentication, protection of personal data, and the establishment of technical standards.

The two documents are closely related. Both call for AI stakeholders to make some of the same overarching ethical commitments: respect for the rule of law and human rights, the use of AI for good, and, in general, ensuring transparency, explainability, safety, security, and overall responsibility in the development and deployment of AI technologies. Each enumerates a nearly identical list of non-binding principles (*see Appendix X*) with the exception that the *Guiding Principles* have one additional principle that the *Code of Conduct* lacks (the 12th: "Promote and contribute to trustworthy and responsible use of Advanced AI systems"). Importantly, neither of the documents is legally binding. They were created with the expectation that governments would develop more detailed and enduring regulations to supplant these voluntary measures.

There are, however, differences. The *Guiding Principles* provide a comprehensive framework applicable to all stakeholders, offering non-binding guidelines to point organizations and governments toward best practices. It is conceived as a "living document" intended to evolve in response to technological advancements. The *Guiding Principles* include, for instance:

- using red teaming and other external testing measures to reliably develop an AI model,
- identifying and mitigating risks post-deployment,
- sharing information with other stakeholders, and
- developing and disclosing a governance and risk management framework that includes privacy policies.

The *Code of Conduct* offers voluntary guidance for "organizations developing the most advanced AI systems, including the most advanced foundation models and generative AI systems,"[2179] without offering specific definitions for these technologies. It builds upon and extends the *Guiding Principles* by formulating a more detailed list of 11 actions that AI organizations should undertake at every stage of an advanced AI's lifecycle, offering details not covered in the *Guiding Principles*. For instance, the 11 principles of the *Code of Conduct* closely align with 11 of the 12 *Guiding Principles*, urging AI actors to proactively identify, evaluate, and mitigate risks. However, while the *Guiding Principles* do *not* specify what those "risks" might include, the *Code of Conduct does*, detailing concerns such as offensive cyber capabilities, weapons development, health, critical infrastructure, and democratic rights. Likewise, the third principle in the *Code of Conduct* calls on AI organizations to ensure transparency through regular reporting to the public. However, only the *Code of Conduct* specifies the content of such technical documentation, including details on the AI system's effects and risks, red-teaming results, and system performance capabilities. The *Code of Conduct* will undergo periodic reviews and updates through regular multi-stakeholder consultations to ensure the proposed measures remain effective and responsive to the rapid advancements in technology.

[2179] *Id.*

6.3.2. Impact of the *Hiroshima AI Process Comprehensive Policy Framework*

The *Hiroshima AI Process Comprehensive Policy Framework* is not a treaty. It is a non-binding framework whose implementation is voluntary for each of the participating G7 countries.[2180] The *Code of Conduct* states that it is incumbent upon countries that choose to implement it to enforce it within their respective jurisdictions.[2181] AI actors and organizations, including members from academia, civil society, and the private and public sectors, may also endorse the *Code of Conduct*. The G7 has urged organizations developing advanced AI to commit to the *Code of Conduct* immediately following the announcement of their agreement to the overall *Framework*.[2182] There is no current list of organizations that have committed to implementing the *Code of Conduct*, though several companies, including Anthropic,[2183] Inflection,[2184] and Milestone Systems,[2185] have expressed their public support. After 2024, the G7 plans to release a list of all organizations that have formally committed to the *Code of Conduct.*[2186]

Despite the possibility of endorsing the *Code of Conduct*, there is no monitoring mechanism in place to ensure and enforce compliance. Some have criticized the document as insufficient and lacking details,[2187] which could make the document even harder to enforce. The G7 has undertaken measures to build a monitoring capability and has been working closely with the OECD to draw on its expertise in soft law implementation. During the June 13-15, 2024, summit of the G7, held in Apulia, Italy, the G7 announced the development of a *reporting* framework developed alongside the OECD to monitor implementation of the *Code of Conduct.*[2188] A pilot of the program is expected during the October session of the Industry, Tech, and Digital Ministers meeting.[2189] In the document's current form, however, companies may signal their support for it, but there is no international mechanism in place to ensure compliance. It is left to

---

[2180] White & Case LLP, *AI Watch: Global regulatory tracker - G7* (May 13, 2024), https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-g7.

[2181] *Hiroshima Process International Code of Conduct for Advanced AI Systems* (Oct. 30, 2023), https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems

[2182] G7 Leaders' Statement, *supra* note 2174.

[2183] Anthropic, *Thoughts on the US Executive Order, G7 Code of Conduct, and Bletchley Park Summit* (Nov. 30, 2023), ANTHROPIC https://www.anthropic.com/news/policy-recap-q4-2023.

[2184] Inflection, *We welcome the G7 Hiroshima Code of Conduct for developing advanced AI systems*, INFLECTION.AI, https://inflection.ai/g7-hiroshima-code-of-conduct (last visited June 20, 2024).

[2185] Milestone, *Milestone Systems First to Adopt G7 Code of Conduct for Artificial Intelligence* (Jan. 18, 2024) https://www.milestonesys.com/company/news/press-releases/first-to-adopt-g7-code/.

[2186] Hiroshima AI Process, SUPPORTERS, https://www.soumu.go.jp/hiroshimaaiprocess/en/supporters.html (last visited June 20, 2024).

[2187] Enza Iannopollo, *The G7 AI Guidelines: Long On Good Intentions, Short On Detail And Substance*, FORRESTER (Nov. 8, 2023), https://www.forrester.com/blogs/the-g7-ai-guidelines-long-on-good-intentions-short-on-detail-and-substance/.

[2188] Apulia G7 Leaders Communiqué (2024), https://www.g7italy.it/wp-content/uploads/Apulia-G7-Leaders-Communique.pdf

[2189] *Id.*

individual state governments to rigorously enforce the spirit of each of the *Code of Conduct*'s 11 action items.

Meanwhile some G7 countries have already adopted measures aligning with the *Framework*. Notably, on July 21, 2023, the Biden Administration announced that leading AI companies in the US had signed commitments agreeing to action items similar to those prescribed in the *Code of Conduct* (*see above section 5.3.2.B(2)*).[2190] On October 30, 2023, the same day the G7 announced the *Code of Conduct*, the White House issued its AI Executive Order.[2191] This order includes several measures outlined in the *Code of Conduct*, such as requirements for federal agencies to show red-teaming results[2192] and to institute threshold risk-management practices.[2193] In Europe, Ursula von der Leyen, President of the European Commission, noted that the *Code of Conduct* would complement the EU AI Act's legally binding rules.[2194] Other countries, such as Canada (*see above section 5.4.2.B*), have also implemented their own voluntary code of conduct with the expectation that legal obligations under future legislation will supplant the code of conduct.[2195] Of course, Japan, too, created its own non-binding soft law to govern AI firms through its *AI Guidelines for Business* (*see above section 5.4.5*).[2196]

6.3.3. Other recent developments

The G7's Industry, Tech, and Digital Ministers met March 14-15, 2024, to discuss global AI regulations.[2197] The group invited Brazil, South Korea, Ukraine, and the United Arab Emirates to join the conference, as well as representatives from the OECD, the United Nations Development Programme (UNDP), UNESCO, the International Telecommunication Union (ITU), and the UN Secretary-General's Envoy on Technology. Those in attendance agreed to update the *Code of Conduct* and Guiding Principles" according to new AI developments.[2198]

[2190] Fact Sheet, White House, Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI (July 21, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

[2191] Exec. Order No. 14110, *supra* note 1527.

[2192] Exec. Order No. 14110, *supra* note 1527.

[2193] *Id.* at § 10.1 IV

[2194] Press Release, European Commission, *Commission welcomes G7 leaders' agreement on Guiding Principles and a Code of Conduct on Artificial Intelligence* (Nov. 1, 2023), https://digital-strategy.ec.europa.eu/en/news/commission-welcomes-g7-leaders-agreement-guiding-principles-and-code-conduct-artificial.

[2195] Gov't of Canada, Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems (Sept. 2023) https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems

[2196] Ministry of Internal Affairs and Communications of Japan & Ministry of Economy, Trade and Industry, *AI Guidelines for Business (draft)* (Jan. 2024), https://www.soumu.go.jp/main_content/000923717.pdf.

[2197] G7 INDUSTRY, TECHNOLOGY AND DIGITAL MINISTERIAL MEETING (Mar. 14–15, 2024), https://assets.innovazione.gov.it/1710505409-final-version_declaration.pdf.

[2198] *Id.* at Annex 3.

Another outcome from the ministerial meetings was the announcement of several G7 research projects that would create tools to improve collaboration and interoperability. These include, for instance, an "AI Toolkit" for assessing the relative advantages and disadvantages of integrating AI within the public sector.[2199] They also include a *Compendium of Digital Government Services*, a collection of examples of when G7 governments have successfully digitized government services; and a *Mapping Exercise of Digital Identity Approaches*, which will chart similarities across G7 countries' digital policies.[2200]

On May 2, 2024, nearly a year after the launch of the *Hiroshima AI Process Comprehensive Policy Framework*, Japanese Prime Minister Kishida Fumio announced the Hiroshima AI Friends Group, a collection of countries who support the spirit of the *Framework* and its voluntary guidelines.[2201] As of May 2024, 49 countries are considered "Friends" of the group.[2202] The announcement, which came during the 2024 annual OECD meeting chaired by Japan, signaled the overall *Framework*'s broader political importance. Even while implementation of the *Framework* remains unclear, it has, nonetheless, become an important step in supporting global collaboration toward generative AI regulation.

6.4. The G20

The Group of 20 —better known as the G20— is composed of the European Union, the African Union, and 19 countries that meet regularly to strategize on issues involving macroeconomic policy, global trade, climate, and health. Together, member countries and regions make for 85% of global economic output and 80% of the world's population.[2203]

6.4.1. The 2019 Osaka summit and the release of the *G20 AI Principles* (2019)

The G20, during its June 2019 Osaka Summit, agreed to a general framework for AI governance, called the *G20 AI Principles*, which is grounded in the OECD Recommendation on Artificial Intelligence. The principles call on AI users and developers to be fair, responsible, and transparent.[2204] The document is brief, with the enumeration of the same five principles the OECD prioritizes:[2205]

[2199] *Id.*

[2200] *Id.*

[2201] Hiroshima AI Process, SUPPORTERS, *supra* note 2186.

[2202] *Id.*

[2203] James McBride et al., *What Does the G20 Do?*, CFR BACKGROUNDER (Oct. 11, 2023), https://www.cfr.org/backgrounder/what-does-g20-do#chapter-title-0-3.

[2204] Masumi Koizumi, *G20 ministers agree on guiding principles for using artificial intelligence*, JAPAN TIMES (June 8, 2019), https://www.japantimes.co.jp/news/2019/06/08/business/g20-ministers-kick-talks-trade-digital-economy-ibaraki-prefecture/.

[2205] *Id.*

1. Inclusive growth, sustainable development, and well being
2. Human-centered values and fairness
3. Transparency and explainability
4. Robustness, security, and safety
5. Accountability

Similar to the OECD Recommendation, the *G20 AI Principles* recommend national policies and recommendations for countries to implement in promotion of these five principles. The recommended government actions range from investing in AI research to preparing workers for a potential major impact on the labor market.[2206]

The *Principles* were released as annexes to a document entitled the "G20 Ministerial Statement on Trade and Digital Economy," which was published following the 2019 G20 Osaka Summit.[2207] There, the G20 reaffirmed its commitment to human-centered AI and to promote governance that ensures AI's responsible development and governance that is agile, flexible, and "innovation-friendly."[2208] The "Ministerial Statement" also underscores the role of civil society and international dialogue in building a responsible digital economy and addressing AI's challenges.

6.4.2. Subsequent G20 summits

The G20 has held several annual summits since 2019. In each, artificial intelligence has been on the agenda to varying degrees. At the 2021 Rome Summit, the G20 emphasized the need to bridge the digital divide and to leverage AI to enhance global healthcare and economic disparities, especially in the wake of the COVID pandemic.[2209] And at the 2022 Bali Summit, the G20 discussed workforce adaptations to potential labor market disruptions.[2210] Most recently, at the 2023 New Delhi Summit, the G20 underscored AI's significant role in advancing sustainable development and reducing inequality. The Summit, hosted by India, also saw Indian Prime Minister Narendra Modi propose the establishment of a global framework for responsible AI governance.[2211] In the same speech before the G20, Modi argued that a good first step toward responsible AI governance would be to guarantee that both data and algorithms are transparent and unbiased.[2212] At the end of the summit, the G20 included an "AI Provision" in its "G20 New Delhi Leaders' Declaration" that reaffirmed the forum's commitment to its own principles.

[2206] *Id.*

[2207] G20 MINISTERIAL STATEMENT ON TRADE AND DIGITAL ECONOMY, https://wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf.

[2208] *Id.*

[2209] Center for AI and Digital Policy, *G20 and Artificial Intelligence*, https://www.caidp.org/resources/g20/, (last visited Aug. 3, 2024).

[2210] *Id.*

[2211] *PM Modi calls for global framework for ethical use of AI*, ECON. TIMES (Dec. 12, 2023), https://economictimes.indiatimes.com/news/india/pm-modi-calls-for-global-framework-for-ethical-use-of-ai/articleshow/105939251.cms?from=mdr.

[2212] *Id.*

The G20's next summit will take place in Rio De Janeiro on November 18-19, 2024,[2213] where it is expected leaders will continue discussion on the *AI Principles* and the ways in which G20 can capitalize on AI's potential for good. Ahead of that annual summit, Brazil, under the G20's Digital Economy Working Group, hosted a conference on "Artificial Intelligence for Sustainable Development and Reduction of Inequality" in April 2024 to address sustainability and other social impacts of AI.

6.5. BRICS

BRICS is an intergovernmental organization chaired by five countries, the first letters of each forming the eponymous acronym (Brazil, Russia, India, China, and South Africa). Initially comprising Brazil, Russia, India, and China, the BRICS group formally expanded in 2011 to include South Africa. In 2024, membership extended to include Egypt, Ethiopia, Iran, and the United Arab Emirates. (Saudi Arabia is said to be considering an invitation, too.[2214])

Over the last decade, AI has occasionally come up during the organization's annual summits.[2215] The BRICS group laid the groundwork for AI cooperation in its 2015 "Memorandum of Understanding on Science, Technology, and Innovation," highlighting information and communications technology as a pivotal area for collaboration.[2216] The BRICS leaders first explicitly mentioned AI in their joint declaration in 2017, identifying it as a policy area where the BRICS countries should enhance their cooperative efforts.[2217] Since the 2017 joint declaration, AI has become a recurring topic in the BRICS group's ministerial meetings. BRICS predominantly views AI as a catalyst for economic growth, development, technological progress, and inclusive societies.[2218] Collaborative efforts and progress in AI are frequently highlighted as essential components of BRICS' AI discourse.

In 2023, the BRICS summit in Johannesburg marked a significant step with the formation of an "AI Study Group" under the BRICS Institute of Future Networks.[2219] The creation of the Study Group was announced

---

[2213] Center for AI and Digital Policy (CAIDP), *G20 and Artificial Intelligence*, https://www.caidp.org/resources/g20/ (last visited June 29, 2024).
[2214] Reuters, *Saudi Arabia has not yet joined BRICS - Saudi official source*, REUTERS (Feb. 1, 2024), https://www.reuters.com/world/saudi-arabia-has-not-yet-joined-brics-saudi-official-source-2024-02-01/.
[2215] Laura Mahrenbach & Mihaela Papa, *BRICS Wants to Shape Global AI Governance, Too*, TUFTS U. FLETCHER RUSSIA AND EURASIA PROGRAM (Mar. 27, 2024), https://sites.tufts.edu/fletcherrussia/brics-wants-to-shape-global-ai-governance-too/.
[2216] Memorandum of Understanding on Cooperation in Science, Technology, and Innovation between the Governments of the Federative Republic of Brazil, the Russian Federation, the Republic of India, the People's Republic of China and the Republic of South Africa, http://www.brics.utoronto.ca/docs/BRICS%20STI%20MoU%20ENGLISH.pdf (last visited June 29, 2024).
[2217] BRICS Leaders Xiamen Declaration, U. OF TORONTO (Sept. 4, 2017), http://www.brics.utoronto.ca/docs/170904-xiamen.html.
[2218] Mahrenbach & Papa, *supra* note 2215.
[2219] GIP Digital Watch Observatory, "BRICS announces formation of AI study group," (August 23, 2023), https://dig.watch/updates/brics-members-announce-formation-of-ai-study-group

by the Chinese President Xi Jingping.[2220] It aims to monitor AI advancements, foster innovation, and establish a robust AI governance framework. President Xi noted that the Study Group would also "develop AI governance frameworks and standards with broad-based consensus."[2221] The Study Group will facilitate cooperation and the exchange of information among BRICS members. However, details remain vague. What is known is that the Study Group will be housed within the BRICS Institute of Future Networks.[2222] And it comes at a time of other AI initiatives, including a Digital Economy Working Group and investments in AI applications by the BRICS-led New Development Bank.[2223] Some are eager to see what kind of regulatory consensus may be achieved now that BRICS has grown to include other countries with competing regulatory approaches.[2224]

These initiatives reflect the BRICS countries' commitment to advancing their AI capabilities and also shaping global AI governance to include diverse perspectives and ensure equitable development across different regions.

6.6. African Union

The African Union (AU) is a continental, intergovernmental organization consisting of 55 member states that comprise the African continent. The AU was officially founded in 2001 and was modeled after the European Union.[2225] Similar to the EU, the African Union includes several important decision-making institutions, though it primarily acts as a forum for discussing regional policies.[2226] Currently, a significant objective for the AU is the fulfillment of its "Agenda 2063," a 50-year development strategy to prioritize social and economic development, wider continental integration, and overall peace and security.[2227]

The AU has taken several steps toward the development of a responsible AI strategy. On February 29, 2024, the African Union Development Agency-New Partnership for Africa's Development (AUDA-NEPAD), which is the AU's technical agency responsible for implementation of "Agenda 2063," published a white paper during its "AI Dialogue" conference.[2228] The paper, titled "Regulation and Responsible Adoption of

---

[2220] Admire Moyo, *BRICS bloc commits to secure, equitable artificial intelligence*, ITWEB (Aug. 25, 2023) https://www.itweb.co.za/article/brics-bloc-commits-to-secure-equitable-artificial-intelligence/mQwkoq6YpLzM3r9A.

[2221] Ministry of Foreign Affairs of the People's Republic of China, *Global AI Governance Initiative* (Oct. 20, 2023), https://www.mfa.gov.cn/eng/wjdt_665385/2649_665393/202310/t20231020_11164834.html.

[2222] *Id.*

[2223] Mahrenbach & Papa, *supra* note 2215.

[2224] *Id.*

[2225] African Union, *About the African Union*, https://au.int/en/overview (last visited June 29, 2024).

[2226] Encyclopaedia Britannica, *African Union* (June 7, 2024), https://www.britannica.com/topic/African-Union.

[2227] African Union, *Agenda 2063: The Africa We Want*, https://au.int/en/agenda2063/overview (last visited June 29, 2024).

[2228] Council of Europe, *Presentation of the Council of Europe's activities on Artificial Intelligence (AI) during the OECD - African Union AI Dialogue* (Mar. 5–6, 2024), https://www.coe.int/en/web/artificial-intelligence/-/presentation-of-the-council-of-europe-s-activities-on-artificial-intelligence-ai-during-the-oecd-african-union-ai-dialogue.

AI in Africa Towards Achievement of AU Agenda 2063,"[2229] was the culmination of two years' work and was developed in collaboration with the AU's High-Level Panel on Emerging Technologies (APET). It stands at over 200 pages long and analyzes each one of five different pillars that are seen as critical in the responsible development of AI adoption. Those five pillars include: human capital development, a strong but responsible data infrastructure, an enabling environment for AI development and deployment, a robust AI economy, and a network of sustainable partnerships.[2230]

The paper delves into each pillar and studies existing strategies while offering recommendations for the further implementation of each pillar. The paper does not address the specific regulatory and legal challenges of generative AI. It does, however, urge any future African Union AI strategy to incorporate ethical principles in the governance and regulation of AI and that African countries, in general, should emphasize legal tools that will enhance values of fairness, safety, privacy, and security.[2231] The document cites the UNESCO guidelines (*see below section 6.11*), in particular, as a model for ethical guidance of responsible AI.[2232] In general, the document seeks to balance these key ethical concerns with creating an enabling environment for the AI industry in Africa. The white paper is seen as a first step toward a wider continental strategy on AI. Such a strategy is expected to dive deeper into the regulatory strategies African countries should implement to enable the trustworthy development of AI.[2233]

In the beginning of February 2024, the AU convened members in Addis Ababa, Ethiopia, for the 44th Ordinary Session of the Executive Council.[2234] At the convention's conclusion, the AU Commission called for the expedited development of a Continental AI Strategy, a comprehensive roadmap for African nations to responsibly develop AI technologies. The Commission tasked a specific Working Group on AI with the task of developing the continental strategy,[2235] which was expected to draw on the lessons and policy recommendations from the AUDA-NEPAD white paper.[2236] From June 11-13, 2024, the AU held the Second Extraordinary session of the Specialized Technical Committee on Communication and information and communications technology.[2237] It was at this session of over 130 African ministers and experts that both the Continental AI Strategy and the African Digital Compact, a separate document detailing Africa's

---

[2229] Center for Strategic & International Studies, *The African Union AI Continental Strategy: Examining the African AI Landscape*, YouTube (Apr. 25, 2024), https://www.youtube.com/watch?v=YwRGvyY6y_A.

[2230] African Union Development Agency, *Regulation and Responsible Adoption of AI in Africa Towards Achievement of AU Agenda 2063*, Auda-Nepad (June 2023), https://dig.watch/resource/auda-nepad-white-paper-regulation-and-responsible-adoption-of-ai-in-africa-towards-achievement-of-au-agenda-2063.

[2231] *Id.*

[2232] *Id.*

[2233] Center for Strategic & International Studies, *supra* note 2229.

[2234] Press Release, *Pan-African Parliament, 44th Ordinary Session of the Executive Council opens* (Feb. 14, 2024), https://pap.au.int/en/news/press-releases/2024-02-14/44th-ordinary-session-executive-council-opens.

[2235] African Union, *Multistakeholder Consultative Sessions on the Development of a Continental Strategy on Artificial Intelligence (AI)* (Apr. 19–24, 2024), https://au.int/en/newsevents/20240419/multistakeholder-consultative-sessions-development-continental-strategy#:~:text=Building%20on%20the%20AU%20AI,regional%20and%20international%20cooperation%20and.

[2236] Center for Strategic & International Studies, *supra* note 2229.

[2237] Press Release, African Union, *African Ministers Adopt Landmark Continental Artificial Intelligence Strategy, African Digital Compact to drive Africa's Development and Inclusive Growth* (June 17, 2024), https://au.int/en/pressreleases/20240617/african-ministers-adopt-landmark-continental-artificial-intelligence-strategy.

strategy to manage its digital future and promote overall societal progress, were introduced and unanimously endorsed.[2238] The AU Executive Council will review the documents for consideration and formal adoption in July 2024.[2239]

6.7. AI safety summits

Since 2023, several significant AI safety summits have taken place, focusing on the challenges and risks associated with AI technologies. The inaugural AI Safety Summit was hosted by the UK at Bletchley Park on November 1-2, 2023, at the behest of the UK Prime Minister. This summit convened representatives from 28 countries, including the United States, China, and the European Union, to address the global challenges and opportunities presented by advanced AI systems. The following AI safety summit took place virtually on May 21-22, 2024, and was hosted by both the United Kingdom and South Korea. France will host the next full in-person summit in February 2025.

6.7.1. The UK AI Safety Summit (November 2023)

The AI Safety Summit hosted by the UK focused on five primary objectives:[2240]

1. achieving a consensus on the risks associated with frontier AI;
2. advancing international cooperation through national and international frameworks;
3. determining suitable safety measures for private sector entities;
4. identifying areas for collaborative safety research; and
5. highlighting beneficial applications of AI.

The Summit resulted in the Bletchley Declaration, which emphasized the urgent need for international collaboration to manage the potential risks associated with advanced AI systems.

6.7.1.A. The summit roundtables

During the Summit, discussions and debates addressed various aspects of AI safety, including potential risks and ethical concerns. The summaries of the roundtable discussions were published.[2241] Some roundtables addressed various risks associated with frontier AI. The first roundtable focused on global

[2238] *Id.*

[2239] *Id.*

[2240] Government of the United Kingdom, *About the AI Safety Summit 2023*, https://www.gov.uk/government/topical-events/ai-safety-summit-2023/about

[2241] Government of the United Kingdom, *AI Safety Summit 2023: Roundtable Chairs' Summaries, 1 November* (Nov. 1, 2023), https://www.gov.uk/government/publications/ai-safety-summit-1-november-roundtable-chairs-summaries/ai-safety-summit-2023-roundtable-chairs-summaries-1-november--2#roundtable-1-risks-to-global-safety-from-frontier-ai-misuse (day 1 summaries); *AI Safety Summit 2023: Roundtable Chairs' Summaries, 2 November* (Nov. 3, 2023), https://www.gov.uk/government/publications/ai-safety-summit-2023-roundtable-chairs-summaries-2-november/ai-safety-summit-2023-roundtable-chairs-summaries-2-november#roundtable-priorities-for-international-attention-on-ai-over-the-next-5-years-to-2028 (day 2 summaries).

safety threats, such as biosecurity and cybersecurity, and it called for urgent cross-sector collaboration. The second roundtable discussed the unpredictability of scaling AI capabilities, highlighting the benefits for healthcare but also the substantial risks and emphasizing the need for rigorous safety testing and monitoring. The third roundtable explored potential existential risks from losing control over advanced AI, advocating for thorough safety testing and further research. The fourth roundtable addressed societal risks, including threats to democracy and human rights. It recommended involving the public in research efforts. The fifth roundtable stressed the need for rapid development of AI safety policies and the importance of governmental regulation, noting that company policies alone are insufficient.

Other roundtables discussed the roles and actions needed from various stakeholders to address AI risks and opportunities. Roundtable 6 emphasized the need for national policymakers to balance risks and opportunities through rapid, agile, and innovative governance, while promoting international collaboration despite differing national contexts. Roundtable 7 focused on the international community's priorities, including developing a shared understanding of AI capabilities and risks, coordinating safety research, and ensuring the widespread benefits of AI. Roundtable 8 highlighted the importance of the scientific community understanding existing risks, collaborating with governments and the public, and avoiding power concentration. Roundtable 9 stressed the need for public skills development and enhancing governmental technical capabilities to maximize AI's potential benefits. Finally, roundtables 10 and 11 centered on future international collaboration, particularly in combating AI-powered disinformation and deepfakes and ensuring that all regions benefit from AI's transformative potential.

These discussions resulted in a plan to establish an international panel of experts who will compile an annual report aimed at initiating worldwide discussions on AI policy and regulation.[2242] The panel will follow a format akin to that used by the International Panel on Climate Change (IPCC) for climate change assessments. This was not the first proposal for an IPCC-like panel for AI. The month prior, in October 2023, Eric Schmidt, former CEO of Google, and several other leading industry executives likewise proposed an IPCC-like panel of experts for AI.[2243]

6.7.1.B. The Bletchley Declaration (November 2023)

At the end of the Summit, 28 attending countries, including the United States, Saudi Arabia, China, the United Kingdom, and the European Union endorsed the "Bletchley Declaration," heralded by the UK government as a pioneering global agreement.[2244] The Bletchey Declaration expresses a collective commitment to proactively manage the potential risks associated with "frontier AI," which refers to highly capable general-purpose AI models (*see above section 2.1.2.A(4)*). The signatories of the Bletchley

[2242] Prime Minister Rishi Sunak, *Prime Minister's Speech at the AI Safety Summit: 2 November 2023*, (November 2, 2023), https://www.gov.uk/government/speeches/prime-ministers-speech-at-the-ai-safety-summit-2-november-2023
[2243] Mustafa Suleyman et al., *Proposal for an International Panel on Artificial Intelligence (AI) Safety (IPAIS)*, CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE, (October 27, 2023), https://carnegieendowment.org/posts/2023/10/proposal-for-an-international-panel-on-artificial-intelligence-ai-safety-ipais-summary?lang=en
[2244] Government of the United Kingdom, *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023* (Nov. 1, 2023), https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

Declaration committed to identifying AI safety risks through rigorous scientific and evidence-based research. They aim to develop risk-based policies to ensure the safe and responsible development and deployment of AI models. The Declaration advocates for the establishment of an internationally inclusive network focused on AI safety research to create evidence-based strategies for managing these risks. This initiative requires collaboration between governments and AI companies to integrate safety measures into AI development processes. While recognizing that different approaches may be taken to achieve these objectives, the Declaration underscores the critical importance of international cooperation.

6.7.1.C. Policy paper on AI safety testing (November 2023)

In addition to the Bletchley Declaration, the Summit produced a policy paper on AI safety testing.'[2245] This document, signed by 10 countries —including the UK, US, and major European states— as well as leading technology companies, establishes a comprehensive framework for testing next-generation AI models by government agencies.[2246] Signatories include leading AI firms, such as OpenAI, Google, Anthropic, Amazon, Mistral, Microsoft, and Meta, as well as government representatives from the US, the UK, Australia, Canada, France, the European Union, Germany, Italy, Japan, South Korea, and Singapore,[2247] who commit to support safety testing of their frontier models.

The document also reiterates that it is incumbent upon governments to evaluate new AI models developed by companies before their public release. This evaluation aims to ensure that these new models do not pose excessive risk to national security.[2248] The agreement encourages international cooperation and supports government agencies in strengthening their capacity for AI testing and developing their own safety regulations. However, it is not legally binding and appears to focus on risk assessments related to national security, rather than addressing potential harms caused by everyday users.

6.7.1.D. International Scientific Report on the Safety Of Advanced AI (May 2024)

The countries represented at the UK AI Safety Summit in November 2023 also agreed to support an independent "State of the Science" report on frontier AI. The *International Scientific Report on the Safety of Advanced AI* was published ahead of the AI Seoul Summit in May 2024.[2249] It provides an updated, science-based assessment of the safety of advanced AI systems.

[2245] Government of the United Kingdom, *Safety Testing: Chair's Statement of Session Outcomes, 2 November 2023* (Nov. 2, 2023), https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-safety-testing-2-november/safety-testing-chairs-statement-of-session-outcomes-2-november-2023.

[2246] *Id.*

[2247] Anna Gross et al., *AI Companies Agree to Government Tests on Their Technology to Assess National Security Risks*, FINANCIAL TIMES (Nov. 2, 2023) https://www.ft.com/content/8bfaa500-feee-477b-bea3-84d0ff82a0de.

[2248] Kiran Stacey & Dan Milmo, *The Great Powers Signed up to Sunak's AI Summit – While Jostling for Position*, THE GUARDIAN (Nov. 2, 2023), https://www.theguardian.com/technology/2023/nov/02/the-great-powers-signed-up-to-sunaks-ai-summit-while-jostling-for-position.

[2249] Bengio et al., *International Scientific Report*, *supra* note 7.

The report highlights several key points about advanced AI models. It emphasizes the potential of AI to enhance public welfare, prosperity, and scientific discovery. The capabilities of AI are advancing rapidly, though the progress on fundamental challenges like causal reasoning remains a matter of debate among researchers. A significant concern is the limited understanding of AI's capabilities and inner workings, which needs improvement. The report addresses the dual nature of AI, acknowledging its potential for both benefit and harm. Malicious use of AI could result in large-scale disinformation, influence operations, fraud, and scams, while malfunctioning AI systems might produce biased decisions affecting protected populations based on race, gender, culture, age, and disability.

The report underscores the uncertainty surrounding AI's future, with numerous possible scenarios. There is disagreement among experts about the future pace of AI advancement, with some predicting slow progress and others expecting rapid or extremely rapid development. The report highlights the need for ongoing international collaboration in AI research and knowledge sharing, and it promotes transparency by including diverse views and perspectives and addressing areas of uncertainty, consensus, and dissent.

Future advancements in AI could lead to systemic risks, including labor market disruption and economic inequality. Opinions vary on the potential for AI to cause catastrophic outcomes if humanity loses control over it. Various technical methods, such as benchmarking, red teaming, and auditing of training data, can mitigate some risks, though these methods have limitations and require further refinement.

6.7.2. The AI Seoul Summit

The AI Seoul Summit, held six months after the UK's AI Safety Summit, convened the same countries and many of the same prominent executive leaders and civil society members who participated in the initial event. This fully virtual summit began with the release of the *International Scientific Report on the Safety of Advanced AI*, an interim report compiled by 75 AI experts from 30 countries, the UN, and the EU.[2250] This report synthesizes the current scientific understanding of general-purpose AI and its associated risks.

Although the two-day virtual forum of the AI Seoul Summit may not have achieved as many significant milestones as the November summit, it still yielded several noteworthy outcomes. Japan, South Korea, and Canada announced the establishment of their own AI safety institutes, while the European Union proposed that the European Commission AI Office could serve a similar role as an AI safety institute.[2251] Ten countries and the EU subsequently signed an agreement to form a global network among their respective AI safety institutes that will include sharing knowledge and aligning safety standards.[2252]

[2250] *Id.*

[2251] Interview by Gregory C. Allen & Georgia Adamson, *The AI Seoul Summit*, in CRITICAL QUESTIONS, CENTER FOR STRATEGIX & INT'L STUDIES, (May 23, 2024), https://www.csis.org/analysis/ai-seoul-summit

[2252] Press Release, Dep't for Sci., Innovation and Tech. et al., *Global leaders agree to launch first international network of AI Safety Institutes to boost cooperation of AI* (May 21, 2024), https://www.gov.uk/government/news/global-leaders-agree-to-launch-first-international-network-of-ai-safety-institutes-to-boost-understanding-of-ai.

The UK and South Korea, the Summit's hosts, also secured a commitment from 16 global AI companies to a set of safety outcomes and accountable governance structures.[2253] Closing off the meeting, the UK and South Korea also secured commitments to continue and deepen work on AI safety research, including the formation of risk thresholds for frontier AI models.[2254]

6.8. The Council of Europe's treaty

The Council of Europe is an intergovernmental organization created in 1949 with a human rights mandate. Headquartered in Strasbourg, France, it is distinct from the European Union. It has 46 member states throughout Europe and is dedicated to the establishment of binding and non-binding legal norms focused on three pillars: human rights, democracy, and the rule of law.[2255]

The Council of Europe is known for drafting over 200 international conventions. Some of the more notable treaties under its ambit include the European Convention on Human Rights,[2256] the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data,[2257] and the Budapest Convention on Cybercrime.[2258]As such, the Council has often forged policies to ensure emerging technologies respect fundamental human rights.

6.8.1. Drafting of the Council of Europe's AI treaty

In September 2019, the Council of Europe's Committee of Ministers established the "Ad Hoc Committee on Artificial Intelligence" (CAHAI), an intergovernmental committee with a two-year mandate (2019-2021).[2259] The Committee released a report in December 2021, advocating for continued discussions on

[2253] Press Release, Dep't for Sci., Innovation and Tech. et al., *Historic first as companies spanning North America, Asia, Europe and Middle East agree safety commitments on development of AI,* (May 21, 2024), https://www.gov.uk/government/news/historic-first-as-companies-spanning-north-america-asia-europe-and-middle-east-agree-safety-commitments-on-development-of-ai.

[2254] Press Release, Dep't for Sci., Innovation and Tech. et al., *New commitment to deepen work on severe AI risks concludes AI Seoul Summit,* (May 22, 2024), https://www.gov.uk/government/news/new-commitmentto-deepen-work-on-severe-ai-risks-concludes-ai-seoul-summit.

[2255] *See* Council of Europe, https://www.coe.int/en/web/portal (last visited July 14, 2024).

[2256] Council of Europe, *European Convention on Human Rights*, https://www.echr.coe.int/documents/d/echr/convention_ENG (last visited June 29, 2024).

[2257] Council of Europe, *Convention 108 and Protocols*, https://www.coe.int/en/web/data-protection/convention108-and-protocol (last visited June 29, 2024).

[2258] Council of Europe, *The Budapest Convention (ETS No. 185) and its Protocols*, https://www.coe.int/en/web/cybercrime/the-budapest-convention (last visited June 29, 2024).

[2259] Council of Europe, *CAHAI - Ad hoc Committee on Artificial Intelligence*, https://www.coe.int/en/web/artificial-intelligence/cahai (last visited July 15, 2024).

drafting a human rights AI treaty.[2260] And the report included a list of measures to be incorporated into a new binding instrument.

In January 2022, the newly formed Committee on Artificial Intelligence (CAI) succeeded CAHAI, continuing its predecessor's work.[2261] The Committee's first draft of the convention was distributed to member states and the European Commission for exclusive review in June 2022.[2262] The text was finalized by the Committee on March 14, 2024.[2263] The Council of Europe officially adopted the *Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law*[2264] in Strasbourg on May 17, 2024. This took place during the Council of Europe's annual meeting of the Committee of Ministers, which brings together the ministers for foreign affairs of the 46 Council of Europe member states.

The Council of Europe comprises 46 Member States, including 27 European Union members, as well as Turkey, Ukraine, and the United Kingdom. Besides its Member States, several countries hold "Observer State" status, enabling them to cooperate with the Council of Europe, participate in its Committees (such as the CAI), and become parties to its conventions. The Observer States include Canada, the United States, Japan, Mexico, and the Holy See. Additionally, the Committee of Ministers approved participation requests from Argentina, Australia, Costa Rica, Israel, Peru, and Uruguay, which also participated as Observer States. Various non-state actors, including civil society organizations, companies, and representatives from international organizations and agencies, such as the OECD, UNESCO, the European Union Agency for Fundamental Rights, and the European Data Protection Supervisor, were involved as Observers in the development of the AI convention.

The Council of Europe collaborated closely with the EU as it drafted its AI Act. The two organizations worked in parallel, with significant interaction. This collaboration was facilitated by the fact that all 27 EU Member States are also members of the Council of Europe, along with 19 other European countries. All participants actively engaged by providing comments and text proposals to the draft treaty until the final day of negotiations.

[2260] Emilio de Capitani, *The COE Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law. Is the Council of Europe losing its compass?*, EUROPEAN AREA OF FREEDOM SECURITY & JUSTICE (Mar. 4, 2024), https://free-group.eu/2024/03/04/the-coe-convention-on-artificial-intelligence-human-rights-democracy-and-the-rule-of-law-is-the-council-of-europe-losing-its-compass/.

[2261] Council of Europe, *Committee on Artificial Intelligence*, https://www.coe.int/en/web/artificial-intelligence/cai (last visited July 15, 2024).

[2262] Council of Europe, *Council of Europe's work in progress* (Jan. 2024), https://www.coe.int/en/web/artificial-intelligence/work-in-progress#01EN.

[2263] European Network of National Human Rights Institutions (ENNHRI), *Draft Convention on AI, Human Rights, Democracy and Rule of Law finalised: ENNHRI Raises Concerns* (Mar. 20, 2024), https://ennhri.org/news-and-blog/draft-convention-on-ai-human-rights-democracy-and-rule-of-law-finalised-ennhri-raises-concerns/#:~:text=On%2014%20March%202024%2C%20the,on%20Artificial%20Intelligence%20(CAI).

[2264] Council of Europe, *Council of Europe adopts first international treaty on artificial intelligence* (May 17, 2024), https://www.coe.int/en/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence#:~:text=The%20convention%20is%20the%20outcome,%2C%20the%20Holy%20See%2C%20Israel%2C.
https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f.

6.8.2. Key features of the treaty

The Council of Europe's treaty represents the first-ever international legally binding treaty on artificial intelligence. Unlike the European Union's AI Act, which applies only to EU member states, this treaty has potential global reach, aiming to establish a minimum standard for protecting human rights from risks posed by AI. The *Framework Convention*'s purpose is to ensure the protection of human rights, the rule of law, and democratic standards in the application of artificial intelligence (AI) systems.[2265] It establishes a comprehensive legal framework that encompasses the entire lifecycle of AI systems. An accompanying Explanatory Report[2266] clarifies that more intricate standards may be established through targeted protocols, which could be implemented as amendments to the *Framework Convention*.

The Council of Europe's *Framework Convention* covers "the activities within the lifecycle of artificial intelligence systems that have the potential to interfere with human rights, democracy and the rule of law."[2267] The *Framework Convention* aims to regulate the activities "undertaken by public authorities, or private actors acting on their behalf."[2268] Although the *Framework Convention* does not automatically apply to the private sector, it requires each signatory to address risks and impacts arising from activities conducted by private actors "in a manner conforming with the object and purpose" of the *Framework Convention*. Each signatory needs to submit a declaration at the time of signature on whether they intend to directly apply the principles and obligations of the Treaty to the private sector or to take "other appropriate measures" to comply with the treaty's provisions.[2269] From this perspective, the convention significantly differs in its scope from the EU's AI Act, which provides comprehensive regulations for the development, deployment, and use of AI systems within the EU internal market. However, the *Framework Convention* and the AI Act share a common strategy of addressing the risks posed by innovation. The core principles and key obligations of the *Framework Convention* include a risk-based approach, even though the convention does not categorize AI systems according to risk.

Chapter II of the *Framework Convention* outlines the general obligations to which each signatory must adhere. These include implementing measures to ensure that activities within the AI system lifecycle are compatible with obligations to protect human rights, as enshrined in applicable international and domestic

---

[2265] Jacques Ziller, *The Council of Europe Framework Convention on Artificial Intelligence vs. the EU Regulation: two quite different legal instruments*, CERIDAP (Apr. 29, 2024), https://ceridap.eu/the-council-of-europe-framework-convention-on-artificial-intelligence-vs-the-eu-regulation-two-quite-different-legal-instruments/?lng=en.

[2266] Council of Europe, *Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, https://rm.coe.int/1680afae67 (last visited June 29, 2024).

[2267] Council of Europe, *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, CM(2024)52-final, article 3; Council of Europe, *133rd Session of the Committee of Ministers* (May 17, 2024), https://search.coe.int/cm?i=0900001680afb11f.

[2268] *Id.*

[2269] *Framework Convention*, art. 3(1)(b).

law.[2270] Additionally, signatories must take steps to ensure that AI systems are not used to undermine the integrity, independence, and effectiveness of democratic institutions and processes, including upholding the principle of separation of powers, respecting judicial independence, and ensuring access to justice.[2271]

Chapter III establishes the general principles that each signatory must incorporate into the measures they implement to ensure compliance with the *Framework Convention*. These principles include: "Human dignity and individual autonomy,"[2272] "Transparency and oversight,"[2273] "Accountability and responsibility,"[2274] "Equality and non-discrimination,"[2275] "Privacy and personal data protection;"[2276] "Reliability,"[2277] and "Safe innovation."[2278]

Chapter IV obligates countries to adopt and maintain sufficient and adequate "remedies" and "procedural safeguards" for situations where human rights are at risk of being violated. The treaty requires comprehensive documentation of AI systems and their usage, making this information accessible to affected individuals. This documentation must be detailed enough to allow individuals to challenge decisions made by or based on the AI system and to contest the use of the AI system itself. The treaty also ensures that individuals have the right to lodge complaints with competent authorities. Additionally, signatories must provide effective procedural guarantees, safeguards, and rights to individuals when an AI system significantly impacts their human rights and fundamental freedoms. Finally, the treaty mandates that individuals be notified when they are interacting with an AI system rather than a human being.

Finally, Chapter V requires countries to adopt and maintain measures for identifying, assessing, preventing, and mitigating AI-related risks that could threaten human rights, democracy, or the rule of law. The treaty mandates comprehensive impact assessments to evaluate both actual and potential effects on human rights, democracy, and the rule of law. Following these assessments, it is crucial to establish effective prevention and mitigation measures. Additionally, the treaty provides that governments must assess the need to implement bans or moratoria on certain AI system applications where they consider such uses incompatible with the respect for human rights, the functioning of democracy or the rule of law.

[2270] *Id.* art. 4, (Protection of human rights).
[2271] *Id.* art. 5, (Integrity of democratic processes and respect for the rule of law).
[2272] *Id.* art. 7.
[2273] *Id*. art. 8.
[2274] *Id*. art. 9.
[2275] *Id.* art. 10.
[2276] *Id.* art. 11.
[2277] *Id.* art. 12.
[2278] *Id.* art. 13.

6.8.3. Implementation of the treaty

Unlike EU directives or regulations, which automatically apply to all the EU Member States, Council of Europe treaties are binding only on states that have signed and ratified them. In other words, this *Framework Convention* will obligate only those states that have formally signed and ratified it. The *Framework Convention* will be formally open to signing on September 5, 2024, during a Council of Europe Minister of Justice conference in Vilnius, Lithuania.[2279]

After signing the treaty, each state must undergo its national ratification process, which typically involves obtaining approval from the national legislature or another designated authority. This step ensures the treaty's provisions are incorporated into the state's domestic law. Once a state completes these internal procedures, it formally submits its instrument of ratification to the Secretary General of the Council of Europe. This ratification document signifies the state's consent to be bound by the treaty. The treaty comes into effect for ratifying states once a specified number of states have ratified it. For the *Framework Convention*, Article 30(3) stipulates that five ratifications, acceptances, or approvals are required, with at least three being from Council of Europe member states, in line with the organization's treaty-making practices.

States which are parties to the treaty must then implement its provisions domestically. It remains uncertain whether the treaty will have direct application —meaning it can be invoked before national courts. This issue varies across different legal systems. For direct application to be feasible, the treaty's provisions must be sufficiently precise to be considered self-executing —an issue which may be debated. In the end, parties to the treaty may either choose to abide by the treaty's measures as written or implement comparable measures.

The *Framework Convention* establishes a follow-up mechanism known as the "Conference of the Parties," which comprises official representatives from the signatory states.[2280] This body is responsible for assessing the implementation of the *Framework Convention*'s provisions. Its findings and recommendations will play a crucial role in ensuring compliance with the *Framework Convention* and in maintaining its long-term effectiveness. Additionally, the "Conference of the Parties" facilitates cooperation with relevant stakeholders and conducts public hearings on key aspects of the *Convention*'s implementation, thereby fostering transparency and inclusivity. Article 25 encourages signatories to exchange relevant information among themselves and to assist non-signatory States in aligning with the *Framework Convention*'s requirements, with the aim of becoming parties themselves.

[2279] Council of Europe, *The Framework Convention on Artificial Intelligence* (May 17, 2024), https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence.
[2280] *Id.* art. 23.

6.8.4. Limitations

The Council of Europe treaty has been praised as an impressive accomplishment[2281] and "a much needed effort to establish internationally agreed upon norms and standards for AI systems."[2282] But it has also been criticized. Over the course of the treaty's development, several Member and Observer States (notably the US, UK, Canada, and Japan)[2283] pushed for the exclusion of *private* sector obligations. Despite public outcry from the European Data Protection Supervisor,[2284] the Parliamentary Assembly of the Council of Europe,[2285] and civil society organizations,[2286] the final draft of the treaty strikes a compromise concerning its application to private actors. Parties to the treaty have a choice: either apply the *Framework Convention* to all private actors or address risks and impacts arising from activities conducted by private actors in a manner conforming with the object and purpose of the convention, taking "other measures to comply with the treaty's provisions while fully respecting their international obligations regarding human rights, democracy and the rule of law."[2287] Countries that opt for the latter are reminded of their other human rights obligations under international law, but they are, nonetheless, free to apply the treaty's provisions to the private sector in a manner they see fit.[2288]

There are several other important and controversial exceptions within the treaty. The treaty explicitly notes that "matters relating to national defense do not fall within the scope of this Convention."[2289] It specifies in an earlier paragraph that the *Framework Convention* does not apply to AI activities that relate to the protection of national security[2290] nor does it apply to research and development activities that concern AI

[2281] OneTrust DataGuidance, *International: CIA finalizes framework convention on AI, Human Rights, Democracy and the Rule of Law* (Mar. 15, 2024), https://www.dataguidance.com/news/international-cai-finalizes-framework-convention-ai.

[2282] Ian Barber, *The world's first treaty on AI: our thoughts and the way forward*, GLOBAL PARTNERS DIGITAL (Apr. 2, 2024), https://www.gp-digital.org/the-worlds-first-treaty-on-ai-our-thoughts-and-the-way-forward/.

[2283] Eliza Gkritsi, *Council of Europe AI treaty does not fully define private sector's obligations*, EURACTIV (Mar. 15, 2024), https://www.euractiv.com/section/digital/news/council-of-europe-ai-treaty-does-not-fully-define-private-sectors-obligations/.

[2284] Press Release,, European Data Protection Supervisor, *EDPS statement in view of the 10th and last Plenary Meeting of the Committee on Artificial Intelligence (CAI) of the Council of Europe drafting the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law,* (Mar. 11, 2024), https://www.edps.europa.eu/press-publications/press-news/press-releases/2024/edps-statement-view-10th-and-last-plenary-meeting-committee-artificial-intelligence-cai-council-europe-drafting-framework-convention-artificial_en.

[2285] Council of Europe, *PACE welcomes draft convention on AI and human rights - but regrets it will not fully cover the private sector*, PARLIAMENTARY ASSEMBLY (Apr. 18, 2024), https://pace.coe.int/en/news/9440/pace-welcomes-draft-convention-on-ai-and-human-rights-but-regrets-it-will-not-fully-cover-the-private-sector.

[2286] *Open Letter to Council of Europe AI Convention Negotiators: Do Not Water Down Our Rights,* ALGORITHM WATCH, (Mar. 5, 2024), https://algorithmwatch.org/en/wp-content/uploads/2024/03/Open_letter_AI_Council_of_Europe.pdf.

[2287] Council of Europe, *The Framework Convention on Artificial Intelligence*, see *supra* note 2279.

[2288] Gkritsi, *supra* note 2283.

[2289] *Framework Convention*, art. 3 § 4.

[2290] Gkritsi, *supra* note 2283.

systems not yet available for use.[2291] Nonetheless, countries are obligated to ensure that all activities still respect international law and human rights obligations under other treaties.

Despite the *Framework Convention*'s broad language and limited scope, it will certainly influence the AI regulation strategies adopted by the members of the Council of Europe.[2292] This treaty marks an important first step in an AI global governance framework that is rooted in human rights.[2293]

6.9. The Global Partnership on AI

The Global Partnership on Artificial Intelligence is a multi-stakeholder initiative that was created based on the model of the Intergovernmental Panel on Climate Change (IPCC).[2294] The Global Partnership, like the IPCC, enlists experts from government, academia, industry, civil society, and other institutions to serve as a kind of one-stop international source for AI expert research.

Launched in June 2020, the Global Partnership was jointly proposed in 2018 by France and Canada following a G7 Employment and Innovation Ministerial Meeting.[2295] Each country recognized the need to advance the G7's call for international collaboration by establishing an "international study group."[2296] This group would both develop expertise and disseminate key insights. The Global Partnership on AI today has been a key player in AI research discussions with four separate working groups under its ambit: Responsible AI, Data Governance, Future of Work, and Innovation and Commercialization.[2297] The Global Partnership's success is due in no small part to the OECD, which hosts a dedicated Global Partnership Secretariat.[2298] In total, 29 countries are members of the Global Partnership on AI.[2299]

The Global Partnership hosts annual meetings and publishes an annual report authored by its own Multi-stakeholder Experts Group (MEG). The 2023 MEG report recapped important projects and reaffirmed the partnership's priorities to harness AI for the purposes of advancing a resilient society and to solve major

---

[2291] *Framework Convention*, art. 3 § 3.

[2292] Robert Spano et al., *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, GIBSON DUNN (June 3, 2024), https://www.gibsondunn.com/council-of-europe-framework-convention-on-artificial-intelligence-and-human-rights-democracy-and-rule-of-law/#:~:text=The%20Council%20of%20Europe%20Has,International%20Treaty%20on%20Artificial%20Intelligence.&text=On%20May%2017%2C%202024%2C%20the,(Convention)%5B1%5D.

[2293] *Id.*

[2294] The Intergovernmental Panel on Climate Change (IPCC) is the United Nations body for assessing the science related to climate change. *See* IPCC https://www.ipcc.ch, (last visited June 20, 2024).

[2295] G7 Employment and Innovation Ministerial Meeting, (Mar. 27–28, 2018), https://www.meti.go.jp/policy/trade_policy/G7G8/pdf/2018_G7innovation_en.pdf.Gov't of Canada,

[2296] *Canada-France Statement on Artificial Intelligence* (June 7, 2018), https://www.international.gc.ca/world-monde/international_relations-relations_internationales/europe/2018-06-07-france_ai-ia_france.aspx?lang=eng.

[2297] Global Partnership on Artificial Intelligence, *About GPAI*, https://gpai.ai/about/ (last visited June 29, 2024).

[2298] *Id.*

[2299] *Community*, GPAI, https://www.gpai.ai/community/ (last visited June 29, 2024).

challenges in climate change, human rights, and global health.[2300] The report also mentions the launch of a Global Partnership Academy.[2301] The Academy has been pitched as a worldwide education awareness project. The main aim is to teach the public about the conditions for AI's controlled development and to educate AI specialists about how to reliably and safely deploy an AI system.[2302]

Meanwhile, the Global Partnership has issued a number of other briefs and reports. These include: "Generative AI, Jobs, and Policy Response;"[2303] "AI Foundation Models & Detection Mechanisms;"[2304] and "Scaling Responsible AI Solutions - Challenges and Opportunities."[2305] This and other work was presented at the last Global Partnership summit which was held in New Delhi, India, from December 12-14, 2023.[2306]

6.10. US-EU Trade and Technology Council

The United States and the European Union have adopted dissimilar strategies for AI governance. Where the US predominantly favors non-binding frameworks and voluntary commitments, the EU favors comprehensive and binding AI legislation. Against this backdrop, the EU-US Trade and Technology Council (TTC) (also known as the US-EU TTC) was created in 2021 to help bridge the growing digital regulatory divide between the two and, more importantly, to foster bilateral trade. The EU is America's largest export partner, and the US is the EU's second largest importer of goods.[2307] Overall, the "transatlantic economy" is estimated to be worth $1.1 trillion.[2308] The TTC, then, is principally a forum for discussing issues that could potentially impact the strong trading relationship between the US and EU. In particular, the TTC's goals[2309] are to:

[2300] *Multistakeholder Expert Group Annual Report*, GPAI, (Nov. 2023), https://gpai.ai/projects/2023-MEG-report.pdf.
[2301] *Id.*
[2302] *Id.*
[2303] Matteo Atamoli et al., *Policy Brief: Generative AI, Jobs, and Policy Response*, GPAI (2023), https://gpai.ai/projects/future-of-work/policy-brief-generative-ai-jobs-and-policy-response-innovation-workshop-montreal-2023.pdf.
[2304] *Id.*
[2305] GPAI, *Scaling Responsible AI Solutions: Challenges and Opportunities* (Dec. 2023), https://www.gpai.ai/projects/responsible-ai/RAI05%20-%20Scaling%20Responsible%20AI%20Solutions%20-%20Challenges%20and%20Opportunities.pdf.
[2306] GPAI, *GPAI Ministerial Declaration* (Dec. 13, 2023), https://gpai.ai/2023-GPAI-Ministerial-Declaration.pdf.
[2307] World Economic Forum, *Here's what to know about the EU-US Trade and Technology Council* (Apr. 17, 2024), https://www.weforum.org/agenda/2024/04/eu-us-trade-technology-council-agreement/.
[2308] Cameron F. Kerry, *Small Yards, Big Tents: How to Build Cooperation on Critical International Standards*, BROOKINGS (Mar. 2024), at 22, https://www.brookings.edu/wp-content/uploads/2024/03/GS_03062024_standards-report.pdf.
[2309] International Trade Administration (ITA), *U.S. - EU Trade and Technology Council (TTC)*, US DEPARTMENT OF COMMERCE,

- deepen economic relations by addressing trade barriers and promoting fair competition,
- secure and diversify supply chains,
- harmonize standards and regulations for emerging technologies, such as AI and cybersecurity,
- uphold shared democratic values, including data protection and the ethical use of technology.

The first TTC ministerial meeting was held September 29, 2021, in Pittsburgh, Pennsylvania[2310] (as of May 2024, there have been six total TTC ministerial meetings[2311]). Since that inaugural meeting, the TTC has spawned a number of initiatives, including several significant steps regarding AI governance and cooperation. In fact, the EU and US have committed to developing three separate projects and initiatives to support AI governance efforts. First, a project that builds on measurement and evaluation tools of trustworthy AI; second, a project that explores and builds upon AI technologies designed to protect privacy; and, third, a joint economic study on AI's potential impact on the workforce.[2312]

6.10.1. The *Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management*

On December 1, 2022, the TTC published a *Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management.*[2313] The document outlines three collaborative projects between the EU and US to align risk policies and build appropriate tools. These include:

- a shared effort to build a common taxonomy and list of terms
- leadership and collaboration to develop international technical standards and tools
- projects that monitor and measure existing and emerging AI risks.

The *Joint Roadmap* calls for the creation of separate working groups to advance each of the three projects, and there have been notable developments within each.

On May 31, 2023, the TTC released the first edition of the *EU-U.S. Terminology and Taxonomy for Artificial Intelligence*. The document listed 65 key AI terms with definitions and references. The first edition of the *Terminology and Taxonomy* was then submitted to external experts to solicit public feedback

https://www.trade.gov/useuttc#:~:text=USEUTTC&text=The%20U.S.%2DEU%20Trade%20and,policies%20in%20shared%20democratic%20values, (last visited July 15, 2024)

[2310] White House, *U.S.-EU Trade and Technology Council Inaugural Joint Statement* (Sept. 29, 2021), https://www.whitehouse.gov/briefing-room/statements-releases/2021/09/29/u-s-eu-trade-and-technology-council-inaugural-joint-statement/.

[2311] European Commission, *Factsheet: EU-US Trade and Technology Council (2021-2024)* (Apr. 15, 2024), https://digital-strategy.ec.europa.eu/en/library/factsheet-eu-us-trade-and-technology-council-2021-2024#:~:text=The%20EU%2DUS%20Trade%20and,based%20on%20these%20shared%20values.

[2312] Alex Engler, *The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment*, BROOKINGS (Apr. 25, 2023), https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/#anchor6.

[2313] European Commission, *TTC Joint Roadmap for Trustworthy AI and Risk Management,* (Dec. 2, 2022), https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management.

from October 27 to November 24, 2023. Gathering the input, the working group updated the *Terminology and Taxonomy* with a revised list of both amended and new terms. On the occasion of the TTC's sixth ministerial meeting, a second edition of the *Terminology and Taxonomy* was created and released on April 5, 2024.[2314] The new document includes changes to 24 of the terms included in the first edition and adds nine new terms.[2315]

The TTC has also been actively involved in the development of international standards for AI. This includes promoting standardization efforts in collaboration with relevant US and EU organizations. In April 2024, during the sixth ministerial meeting in Leuven, Belgium, the EU and US announced the creation of a new dialogue between the EU's AI Office and the US's AI Safety Institute.[2316] This dialogue aims to foster collaboration on AI safety and the ethical use of AI technologies.

The *Joint Roadmap* also mentions the development of "tools for trustworthy AI and risk management." Press releases following ministerial meetings have noted progress on two specific projects related to these tools.[2317] The first is a shared knowledge base of "metrics and methodologies for measuring AI trustworthiness, risk management methods, and related tools."[2318] Second is a set of studies that will look at the saturated landscape of existing standards and tools for the development of trustworthy AI. The hope is that such studies will facilitate the development of coherent standards through the identification of commonalities and gaps among the many existing standards.[2319]

A third project under the *Joint Roadmap* includes other knowledge-sharing mechanisms with the goal of monitoring and measuring potential risks. These include mechanisms to evaluate risks and a tracker that will log existing and emerging risks and categorize these according to context, use cases, and empirical data.[2320] The EU and US have each made a commitment to take actionable steps toward developing these projects.[2321]

6.10.2. Privacy-enhancing technologies

[2314] European Commission, *EU-U.S. Terminology and Taxonomy for Artificial Intelligence - Second Edition*, (Apr. 5, 2024), https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence-second-edition.
[2315] *Id.* at 3.
[2316] Press Release, EU-US Trade and Technology Council (TTC), E*U and US continue strong trade and technology cooperation at a time of global challenges*, (April 5, 2024), https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1827.
[2317] EU-US TTC, *TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management* (Dec. 1, 2022), https://www.nist.gov/system/files/documents/2022/12/04/Joint_TTC_Roadmap_Dec2022_Final.pdf.
[2318] *Id.* at 4.
[2319] *Id.* at 5.
[2320] *Id.*
[2321] *Id.*

Another initiative saw the EU and US embarking on a pilot project focused on Privacy-Enhancing Technologies (PETs).[2322] The TTC initiated pilot projects to assess the use of privacy-enhancing technologies and synthetic data, particularly in the fields of health and medicine. These technologies include federated learning, a machine-learning approach where a model is trained across multiple decentralized devices or servers holding local data samples, without exchanging the data itself. They also include differential privacy, a technique that ensures individual data points in a dataset cannot be identified by adding controlled random noise. Overall, these technologies' aim is to enable extensive data analysis while preserving a degree of data privacy.

In the wake of the third TTC ministerial in December 2022, the EU and the US announced a joint venture to pioneer PETs, specifically for health and medicine applications.[2323] In a subsequent addendum, the two countries agreed to fund joint AI research projects across various other sectors.[2324]

6.10.3. AI's impact on the workforce

The third TTC project, jointly written by the European Commission and the White House Council of Economic Advisors, is a report titled, "The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America."[2325] Published December 5, 2022, the report explores AI's potential impact on the workforce across the EU and US by synthesizing each region's academic research. Overall, the study underscores the numerous risks and opportunities that AI poses to the job market. Further AI integration among businesses will allow firms to scale, lower costs, and make better decisions. At the same time, many jobs will be exposed to AI. Many workers may find their work automated or made fundamentally different because of AI. To mitigate these risks and move AI development in a more positive direction, the report highlights several policy measures. Among them: increased investment in training and job transition services, encouraging investment in AI that helps, rather than hurts, workers, and, finally, investing in regulatory agencies to supervise AI systems and ensure fair hiring and management practices.[2326]

The TTC has itself tried to carry out some of these policy measures through its own initiatives. Following the December 2022 ministerial meeting, the US and EU agreed to create the Talent for Growth Task Force,

---

[2322] U.S. Dep't of Commerce, *FACT SHEET: U.S.-EU Trade and Technology Council Advances Concrete Action on Transatlantic Cooperation* (Dec. 5, 2022), https://www.commerce.gov/news/fact-sheets/2022/12/fact-sheet-us-eu-trade-and-technology-council-advances-concrete-action.

[2323] Raluca Csernatoni, *Towards Strengthening the Transatlantic Tech Diplomacy: Trustworthy AI in the EU-U.S. Trade and Technology Council*, TRANSATLANTIC TECH. AND TRADE FORUM (2023), https://www.transatlantic.org/wp-content/uploads/2023/01/Csernatoni_Background-Paper-on-the-EU-US-TTC-Cooperation-on-AI.pdf?mc_cid=5c3d87eca1.

[2324] *Id.*

[2325] White House, *The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America* (2022), https://www.whitehouse.gov/wp-content/uploads/2022/12/TTC-EC-CEA-AI-Report-12052022-1.pdf

[2326] *Id.*

which would look at ways in which both regions could better train workers and diversify hiring practices.[2327] The taskforce ended its operations at the end of the sixth ministerial meeting. Ahead of the committee's conclusion, its members adopted a final statement, noting that it has influenced private corporations to launch consortiums to better understand the impact AI will have on jobs. The Task Force also discussed and recommended "micro-credentials," skill-based credentials that can be completed within a short time frame, help upskill workers, and meet technical labor needs.[2328]

6.10.4. The AI Code of Conduct

EU Executive Vice President Margrethe Vestager announced plans in May 2023, during the fourth ministerial, for the TTC to draft an "AI Code of Conduct."[2329] The TTC said it would work closely with the G7 and other partners to draft the code.[2330] This announcement, of course, came at the same time the G7 announced the *Hiroshima Process* and was, likewise, taking steps towards a voluntary code of conduct. The two institutions seemingly merged their efforts.[2331] The resulting document was formalized as "the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems," (*see above section 6.3.1*) which speaks to the TTC's success in bridging the transatlantic divide and collaborating with other relevant institutions as well.

6.11. UNESCO

The United Nations Educational, Scientific, and Cultural Organization (UNESCO) is a UN agency that promotes international cooperation and research in education, science, culture, communication, and information. The agency has issued its own AI guidelines. On November 23, 2021, UNESCO adopted the "Recommendation on the Ethics of Artificial Intelligence,"[2332] a text that offers non-binding recommendations for implementing AI ethical principles. In September 2023, it also published a global guidance on generative AI in education and research.

6.11.1. Recommendation on the ethics of AI (2021)

[2327] International Trade Administration, *Talent for Growth Task Force*, U.S. DEPARTMENT OF COMMERCE, https://www.trade.gov/useuttc-taskforce

[2328] *Id.*

[2329] Marianna Drake et al., *EU and US Lawmakers Agree to Draft AI Code of Conduct*, INSIDE PRIVACY (June 9, 2023), https://www.insideprivacy.com/artificial-intelligence/eu-and-us-lawmakers-agree-to-draft-ai-code-of-conduct/#:~:text=On%2031%20May%202023%2C%20at,of%20Conduct%20in%20advance%20of.

[2330] Natasha Lomas, *EU and US lawmakers move to draft AI Code of Conduct Fast*, TECHCRUNCH (MAY 31, 2023), https://techcrunch.com/2023/05/31/ai-code-of-conduct-us-eu-ttc/#:~:text=In%20a%20read%2Dout%20of,how%20to%20ensure%20verification%20(red

[2331] Kerry, *supra* note 2308.

[2332] UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (Nov. 23, 2021), https://unesdoc.unesco.org/ark:/48223/pf0000381137_fre.

The UNESCO Recommendation offers a normative framework for the ethical governance of artificial intelligence through an enumerated list of values, principles, and areas for policy action.[2333] "Values" are "motivating ideals" intended to guide AI governance and shape organizations' and regulators' behavior toward responsible AI development. These "values" include the respect, protection, and promotion of human rights, sustainability, and other vital elements that help advance a more peaceful and interdependent society. "Principles" build upon these values and contextualize them for implementation in the AI context. Those principles, ten total, include some of the following:

- **Proportionality and Do No Harm:** The choice of an AI system and method should be proportionate to the legitimate objective it seeks to accomplish and should not infringe upon an individual's human rights or cause other harms.
- **Fairness and Non-Discrimination:** AI systems must be fair and non-discriminatory according to international law, and AI actors should promote social justice.
- **Human Oversight and Determination:** At any stage of an AI system's life cycle, there must be clear ethical and legal attribution and, in cases of remedy, there must be attribution to a physical person or legal entity. Ultimately, an AI system cannot replace human responsibility and accountability, and a human should never cede decision-making authority to an AI system in a life or death situation.
- **Awareness and Literacy:** Governments, companies, civil society organizations, and other stakeholders should work to promote awareness and education of how AI systems work and the value of data.

Finally, the UNESCO Recommendation suggests 11 areas for policy action. Key suggestions include:

- **Data protection**: The Recommendation seeks to ensure the protection and privacy of personal data to reflect the values of human dignity and autonomy. Personal data must be used, and deleted, in accordance with international laws and the Recommendation's values. The Recommendation also encourages the strengthening of the authority of government regulators to enforce these provisions.
- **Prohibition of "social credit," mass surveillance, and legal personality**: The Recommendation opposes the use of AI for social rating, mass surveillance, and any "highly invasive" technologies that infringe on fundamental freedoms and human rights. It opposes granting AI a legal personality.
- **Support for monitoring and evaluation**: The Recommendation proposes the development of tools to help countries and companies evaluate the impact of AI systems on individuals, society, and the environment. It also encourages member states to assess the status of their legal and technical infrastructures, to introduce frameworks for ethical impact assessments, and to take appropriate measures for adhering to a code of ethics.
- **Environmental protection**: The Recommendation encourages all stakeholders to give priority to the efficient use of data, energy, and other resources, and it calls on AI actors to use AI systems toward solving sustainability issues. The Recommendation discourages the use of systems with a disproportionately negative impact on the environment.

[2333] *Id.*

While the scope of the Recommendation is exhaustive, it does not specify a mechanism for ensuring its effective implementation. At best, the Recommendation suggests the development of a "UNESCO methodology" to gauge the effectiveness of its policies. UNESCO has backed the opening of an International Research Center on Artificial Intelligence (IRCAI) in Slovenia, which will operate under UNESCO's auspices. Recently, IRCAI made news with a published report that found pervasive gender bias across generative AI tools.[2334] It pointed to three responsible sources for the bias: deficient training data, incomprehensive algorithmic selection, and faulty deployment. Most recently, in 2024, UNESCO hosted, in Kranj, Slovenia, the Global Forum on the Ethics of Artificial Intelligence. There, it announced the creation of the Global AI Ethics and Governance Observatory. UNESCO's AI Observatory monitors countries' AI policy landscapes, gathers resources to promote best practices, and promotes the UNESCO Recommendation. The Observatory is a collaborative effort among UNESCO, the Alan Turing Institute, and the International Telecommunication Union.[2335]

6.11.2. Guidance on generative AI in education and research

In September 2023, UNESCO published its first-ever global guidance on generative AI in education and research.[2336] Building on the agency's Recommendation, the guidance looks at how generative AI can be used to support classroom instruction and offers measures and advice for responsibly incorporating the technology into teaching, learning, and research activities.[2337]

The guidance covers various aspects, including an explanation of generative AI technologies, the ethical and policy issues surrounding them, and the steps needed for their regulation. It emphasizes a human-centered approach, advocating for the ethical, safe, equitable, and meaningful use of AI in educational settings. Moreover, it discusses how generative AI can be creatively used in curriculum design, teaching, and learning activities while addressing long-term implications for education and research.

UNESCO has also called on governments to implement appropriate regulations and teacher training to manage the disruptions caused by generative AI technologies. The organization stresses the need for public engagement and governmental safeguards to ensure that AI benefits are maximized while minimizing potential harms and biases.

[2334] Jon Gold, *UNESCO finds 'pervasive' gender bias in generative AI tools*, CIO (Mar. 7, 2024), https://www.cio.com/article/1312311/unesco-finds-pervasive-gender-bias-in-generative-ai-tools.html?amp=1#origin=https%3A%2F%2Fwww.google.com%2F&cap=swipe,education&webview=1&dialog=1&viewport=natural&visibilityState=prerender&prerenderSize=1&viewerUrl=https%3A%2F%2Fwww.google.com%2Famp%2Fs%2Fwww-cio-com.cdn.ampproject.org%2Fc%2Fs%2Fwww.cio.com%2Farticle%2F1312311%2Funesco-finds-pervasive-gender-bias-in-generative-ai-tools.html%3Fusqp=mq331AQIUAKwASCAAgM%25253D&_kit=1.

[2335] Laia Güell Paule, *UNESCO launches Global AI Ethics and Governance Observatory at the 2024 Global Forum on the Ethics of Artificial Intelligence*, DIGITAL SKILLS & JOBS PLATFORM (Feb. 6, 2024), https://digital-skills-jobs.europa.eu/en/latest/news/unesco-launches-global-ai-ethics-and-governance-observatory-2024-global-forum-ethics.

[2336] Fengchun Miao & Wayne Holmes,*Guidance for generative AI in education and research*, UNESCO (Sept. 7, 2023), https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research.

[2337] *Id.*

This initiative reflects UNESCO's commitment to supporting global education systems in navigating the complexities of AI integration and ensuring that these technologies contribute positively to human development.

KEY TAKEAWAYS

- Considering the global deployment of AI, it is crucial to develop and enforce principles on an international scale to prevent and mitigate its most serious risks. The most often-cited dangers are those involving loss of control and the potential for malicious actors or rogue states to misuse the technology to wreak havoc or threaten harm. In this context, various categories of international actions can be identified: drafting international treaties, such as the one adopted by the Council of Europe; guiding member states of supranational organizations in advancing AI regulations, as exemplified by the African Union's initiatives; adopting global recommendations and guidelines, as done by the UN and UNESCO; leading international discussions within diplomatic frameworks like the G7, G20, BRICS, or TTC; supporting state and international policies with recommendations and studies, as exemplified by the OECD; and convening state and industry representatives at AI Safety Summits or through partnerships, such as the Global Partnership on AI.

- The Council of Europe's *Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law* represents the first-ever international legally binding treaty on artificial intelligence. Unlike the European Union's AI Act, which applies only to EU member states, this treaty has potential global reach, aiming to establish a minimum standard for protecting human rights from risks posed by AI. By adopting a risk-based approach to AI systems design, development, use, and decommissioning, it requires thorough consideration of any potential negative consequences these systems might present. While the treaty primarily targets the public sector and private companies acting on its behalf, it offers two options for regulating the private sector. Parties can either directly adhere to the relevant treaty provisions or adopt alternative measures that ensure compliance with the treaty's principles, while fully respecting their international obligations concerning human rights, democracy, and the rule of law.

- Other supranational organizations besides the EU have initiated efforts to guide their member states. Specifically, in February 2024, the African Union published a white paper titled "Regulation and Responsible Adoption of AI in Africa Towards Achievement of AU Agenda 2063." This document analyzes key objectives critical for the responsible development and adoption of AI. More recently, in April 2024, the African Union released the Continental AI Strategy, a comprehensive roadmap for African nations to responsibly develop AI technologies, along with the African Digital Compact, which outlines Africa's strategy for managing its digital future and promoting overall societal progress.

- Major international organizations are also closely monitoring the development of AI and making decisions within their areas of competence. In October 2023, the United Nations established a High-

Level Advisory Body on Artificial Intelligence, a multi-stakeholder group tasked with providing recommendations for international AI governance. The UN General Assembly adopted its first resolution on AI in March 2024, which focuses on "Seizing the opportunities of safe, secure, and trustworthy artificial intelligence systems." Although non-binding, the resolution emphasizes the role of international law in governing AI. Finally, in April 2024, the UN released the zero draft of its forthcoming Global Digital Compact (GDC), which is set to be a governmental, yet non-binding, guide to digital cooperation among UN-led multi-stakeholders. For its part, UNESCO adopted the "Recommendation on the Ethics of Artificial Intelligence," providing non-binding guidelines for the ethical implementation of AI principles on November 23, 2021. Additionally, in September 2023, UNESCO released global guidance on the use of generative AI in education and research.

- The usual frameworks for major diplomatic discussions also serve as privileged venues where states can agree on fundamental principles to guide the global development of AI. From this perspective, the example of the G7 is certainly the most significant. At the Hiroshima Summit in May 2023, the G7 initiated the *Hiroshima AI Process*, a ministerial forum designed for G7 ministers to discuss AI governance and collaborate on developing an international framework. This initiative led to the publication of the *Hiroshima AI Process Comprehensive Policy Framework*, which includes the *Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems* and the *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems.* Although non-binding, these frameworks require AI developers and deployers to implement robust risk management policies and security controls and to emphasize the importance of continuous monitoring, reporting, and mitigating misuse and incidents. The *Hiroshima Process* has proven to be highly influential, both among G7 member states and AI companies. AI has also been a prominent topic in the G20's agenda, with the publication of the G20 AI Principles in 2019. Since then, AI has been discussed at every G20 summit.

- The discussions between the EU and the US within the framework of the EU-US Trade and Technology Council (TTC) are highly significant. The "Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management," published in December 2022, outlined several collaborative AI projects. Following this, the TTC released the "EU-U.S. Terminology and Taxonomy for Artificial Intelligence" and has been actively involved in developing international standards for AI.

- Among major international diplomatic discussions, the case of BRICS warrants special consideration, particularly as artificial intelligence has become a recurring topic in the group's ministerial meetings. BRICS is a five-nation intergovernmental group of Brazil, Russia, India, China, and South Africa. In 2023, it announced the formation of an "AI Study Group" to monitor AI advancements, foster innovation, and establish a robust AI governance framework and standards based on broad consensus. This initiative should be analyzed in the context of China's "Global AI Governance Initiative," introduced by President Xi Jinping in October 2023. This initiative aims to shape the development and governance of AI on a global scale, emphasizing international collaboration and equitable AI governance. It opposes technological monopolies, promotes global cooperation to prevent AI misuse, and underscores the importance of giving developing countries a significant voice in global AI governance.

- As an international research and policy organization, the OECD has produced influential recommendations and studies on AI. In May 2019, the OECD adopted the official "Recommendation on Artificial Intelligence," the world's first intergovernmental standard on AI. Although non-binding, this document was endorsed by all 38 OECD members and eight non-members, including Brazil, Egypt, and Singapore. The Recommendation has served as a foundation for the G20's AI Principles and has significantly influenced the legislative drafting of the European Union's AI Act and other national initiatives. The OECD also established the OECD AI Policy Observatory, which serves as a comprehensive research hub for policymakers and AI experts, while promoting the OECD's guidelines.

- International initiatives often include efforts to bring together government representatives, industry leaders, and AI experts to discuss the risks and challenges of AI, as well as potential measures to address them. One notable example is the Global Partnership on Artificial Intelligence, which aims to be a central international resource for AI expertise, enlisting experts from government, academia, industry, civil society, and other institutions. Furthermore, since 2023, several pivotal AI safety summits have been held to address the challenges and risks associated with AI technologies. The inaugural summit, held in the UK on November 1-2, 2023, brought together representatives from 28 countries, including the United States, China, and the European Union. The following summit took place virtually in May 2024, co-hosted by the United Kingdom and South Korea. France is set to host the next in-person summit in February 2025.

Chapter 7 Final conclusions

The introduction to this report highlighted the regulatory trade-offs and challenges inherent in governing all emerging technologies. In the case of artificial intelligence, these challenges are particularly exacerbated by the uncertainty surrounding AI's potential risks and future trajectory. In this context, the report examined the principal characteristics of generative AI and the associated risks, as commonly referenced and analyzed in contemporary scholarship. It subsequently explored the primary practices employed by generative AI companies to maximize the quality and safety of their tools. Furthermore, the report scrutinized the key regulatory frameworks pertaining to generative AI and major international initiatives in this domain. Several significant conclusions arise from this comprehensive analysis.

1- On the international stage, representatives of leading AI companies have advocated for the adoption of measures at both international and supranational levels (*see section 1.3.2*). They have proposed the establishment of independent, expert-led bodies to objectively inform governments about the current state of AI capabilities and to provide evidence-based predictions about future developments. They have emphasized the necessity of creating a common set of shared standards and principles to guide both nation-states and companies. Finally, industry representatives, along with experts and researchers, have called for the drafting of an international AI treaty to mitigate AI risks and ensure that AI benefits all of humanity.

These calls have been heeded to a certain extent. Over the past year, AI safety summits have convened representatives from numerous countries, resulting in the adoption of the Bletchley Declaration and the publication of the *International Scientific Report on the Safety of Advanced AI*, an interim report compiled by 75 AI experts. The OECD has published numerous influential recommendations and studies on AI. The Global Partnership on Artificial Intelligence serves as an international resource for AI expertise. Various international discussions have led to the proclamation of shared principles, such as those outlined in the G7 *Hiroshima AI Process Comprehensive Policy Framework*. The BRICS-formed "AI Study Group" also aims to establish a robust AI governance framework and standards based on broad-based consensus. Advocates for a true international AI treaty can find satisfaction in the adoption of the Council of Europe's *Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law*, which is the first-ever international legally binding treaty on artificial intelligence.

However, not *all* these calls have been heeded. These initiatives have limitations. The principles discussed at the international level often remain exceedingly broad and are predominantly non-binding. The provisions of the Council of Europe's treaty are also very general and grant significant latitude to the signatory countries in regulating the private sector. Currently, no international binding provisions include limitations on AI model capabilities or global compute thresholds for AI training. The establishment of an international compliance commission to monitor adherence to specific principles is not envisaged. Finally, it is far from certain that an AI system certified as safe in one jurisdiction would be recognized as safe in all others, given that binding regulatory frameworks vary across regions and countries.

2- At the national level, while some industry leaders advocate for regulation, governments encounter a series of complex decisions, each with its own benefits and drawbacks. An aggressive approach to mitigating all hypothetical risks of AI may stifle technological development. Conversely, an overly timid

stance may forfeit the opportunity to guide the industry toward safe technological advancement and away from foreseeable harms. This dilemma is further complicated by the fact that governments often lack the requisite expertise to design and implement a new regulatory framework, while the necessary knowledge and proficiency in AI predominantly reside within the private sector.

Within this context, three distinct strategies were identified in Chapter 1 (*see section 1.2*): self-regulation, co-regulation, and traditional government regulation. Self-regulation allows firms within the industry to adhere to a set of rules or principles they develop. Co-regulation involves having regulators and state agencies collaborate with private actors to develop and implement standards and best practices. And traditional government regulation is where the government enacts laws or binding legal frameworks. Currently, all of these strategies are present among the countries examined in this report, although in practice, they are often mixed.

- Among the countries favoring a laissez-faire approach that allows companies to self-regulate, the case of the United Kingdom is most illustrative. Aware of the challenges and risks associated with AI, the UK government has taken a leading role in organizing international AI safety summits in the UK and South Korea (*see section 6.6*). Additionally, it has established an AI Safety Institute focused on building internal capabilities to assess the safety of advanced AI systems. However, the UK government has thus far advocated for a "wait and see" approach to AI regulation, justifying this stance with the objective of fostering innovation. Currently, the UK relies on existing sector-specific regulatory agencies, some of which have recently updated their regulatory approaches.

  Other countries favor a soft "co-regulation" approach through the publication of non-binding principles and guidelines. Singapore, for instance, has adopted a soft law approach to AI governance, focusing on practical guidance with tools to facilitate implementation. Its *Model AI Governance Framework for Generative AI* aims to guide organizations in developing or deploying generative AI, emphasizing safety, accountability, transparency, and security. Similarly, Saudi Arabia has adopted *AI Ethics Principles*, while the United Arab Emirates does not anticipate enacting an AI law. UAE champions "regulatory sandboxes," which involve live testing of AI in a controlled environment under direct regulatory supervision. In Israel, where the focus has primarily been on promoting the tech sector, potential regulatory changes may introduce legally binding or voluntary standards.

  Until recently, the United States favored self-regulation, allowing AI companies to develop and release increasingly powerful and sophisticated generative AI tools without constraints. Following the release of ChatGPT in the fall of 2022, the Biden administration prioritized dialogue with major AI developers, securing voluntary commitments from leading AI companies in July and September 2023. This approach of "encouraged self-regulation" coexists with federal agencies increasingly assuming a more significant role. Notably, in October 2023, President Biden issued Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The executive order provided guiding principles and policy priorities for the federal administration and agencies. Within this framework, the National Institute of Standards and Technology (NIST) has published various documents establishing non-binding standards developed through joint collaboration with stakeholders.

- At the other end of the spectrum, some jurisdictions have unequivocally restricted AI companies' ability to adopt and implement their own standards. China exemplifies this approach for authoritarian regimes. It has enacted a comprehensive set of laws specifically targeting algorithmic recommendations, deep synthesis technologies, and generative AI services. Some objectives of these laws align with those in democratic regimes —such as the labeling of synthetically generated content to prevent public confusion and the protection of privacy or intellectual property. The primary aim is to prevent illegal activities, such as spreading "fake news" and content that violates "socialist values." Chinese laws impose stringent measures on both service providers and users, including requirements for user identification, activity monitoring, and swift action against illegal content. Generative AI tools must be trained to minimize the production of prohibited content, and those capable of influencing public opinion or mobilizing the public must register with the relevant regulators.

- In this context, Europe appears to occupy a middle ground between the "encouraged self-regulation" approach in the United States and the authoritarian, state-centered approach of China. The necessity for a comprehensive law governing the highest-risk uses of AI became evident in the EU in the early 2020s. Consequently, the AI Act regulates AI applications based on their risk level, which is predominantly sector-based. However, the drafters of the AI Act have broadened their ambitions throughout the negotiation process to also address the technology itself, with many provisions covering General Purpose AI models. While embodying traditional regulatory methods, the European AI Act also incorporates a co-regulation strategy, as its phased implementation will require the adoption of numerous application measures and involve various stakeholders, including AI companies and independent experts.

  The EU's regulatory approach is influencing other countries currently engaged in the process of AI regulation. The soon-to-be-adopted Brazilian AI Act implements a "risk-based approach" based on a gradation of risks. In Canada, the forthcoming Artificial Intelligence and Data Act draws inspiration from various international frameworks, including the AI Act. Some countries that initially refrained from adopting binding AI regulations are gradually moving toward implementing them. The Japanese government is now considering the adoption of a binding legal framework, particularly for high-risk AI systems and those with significant potential impact and risk if misused. Similarly, the Indian government is contemplating the inclusion of provisions to regulate AI systems, especially high-risk ones, in the forthcoming Digital India Act. South Korea is also considering its own "AI Act," which is currently under review by the National Assembly.

3- AI companies have implemented diverse risk mitigation practices as they have faced increasing public scrutiny and evolving risks associated with generative AI. While AI companies have formed industry groups, many current practices stem from individual initiatives by generative AI developers and are sometimes adopted by competitors. These initiatives signify a movement toward genuine self-regulation by establishing industry standards; however, they remain entirely voluntary commitments without independent oversight. Additionally, these practices do not always prioritize self-regulation. For example, practices like red teaming or reinforcement learning primarily focus on technological advancement and enhancing the quality and safety of AI models. Nonetheless, in the medium to long term, current industry initiatives may

contribute to the creation of self-regulatory instruments. They might even be acknowledged by regulators, either as part of non-binding frameworks, such as the NIST frameworks or within formal legal frameworks like the EU AI Act.

Foremost, the rapid pace of technological advancement, characterized by increasingly efficient AI models that succeed one another at an accelerated rate, must be emphasized. This swift evolution and progress impact not only foundation models but also all applications developed based on these models. Consequently, the risks associated with these increasingly powerful models and their applications evolve very quickly. Similarly, developers' practices and methods for risk mitigation are being developed and improved at a very rapid pace. Therefore, it is essential in this context that regulators maintain and foster close ties with the technological community to precisely adapt the standards they wish to uphold in the interest of the greater good. This, of course, necessitates not only effective collaboration between authorities and the industry but also that regulators possess the necessary internal expertise.

4- Several high-level principles and observations emerge from this exploration of the different initiatives related to AI:

- **Regulation of the technology or its applications**: In many countries, sector-specific laws allow AI regulation to evolve incrementally, making narrow adjustments to accommodate changes introduced by AI. However, due to the uncertain reach and implications of AI technology and the development of general-purpose AI models, predicting future applications and use cases is challenging. Consequently, regulating the technology itself is particularly important. This necessity explains why some countries, such as the EU and Brazil, have adopted specific frameworks targeting AI broadly. Additionally, many countries, including the United Kingdom, the United States, Japan, and the European Union, have established dedicated AI safety institutes.

- **The importance of transparency and auditing**. Precisely because the impacts of generative AI are difficult to determine, transparency in the development of this technology is critical. "Model cards" and disclosures about training data represent only the initial steps towards the necessary transparency. To fully understand the implications of foundation models and generative AI applications, both developers and external third parties must rigorously test them prior to deployment. This testing should aim to evaluate performance, biases, alignment, and the potential to generate significant risks.

- **Regulations and enforcement**. Given the complexity and rapid pace of technological development, legislation alone cannot adequately specify rules in advance to govern AI development and applications, even in the near future. Enforcement will be equally, if not more, important than legislation. Effective enforcement will necessitate government recruitment of AI talent, which is both costly and scarce. Additionally, it will require ongoing coordination between the government, companies, and civil society to provide continuous guidance on how existing regulations apply to new and emerging contexts.

- **The relative power of the public and private sectors**. Almost all current generative AI models have been developed by private companies. The need to collect vast amounts of data, overcome the scarcity of chips, and meet the high costs of computation have concentrated the resources required to develop and train the most powerful models in the hands of only a few private companies. To "democratize" the production of AI may require massive public investment to ensure actors other than those tied to the profit-maximizing mission of the firms are able to produce the cutting edge AI models.

- **The promise and risks of open models.** Although private companies are developing the most powerful models, some are publicly releasing the models and their weights. Meta, with its Llama models, leads in producing powerful open models. Other companies, such as Mistral and Falcon, have also published significant open models. These open models promise to democratize the benefits of AI, making them accessible worldwide. Additionally, they could foster a competitive environment quite different than social media and search, which have been controlled by a few oligopolistic actors. However, the openness of these models also raises significant concerns. Once released, they can be used and fine-tuned by malicious actors for various intended and unintended purposes. Furthermore, once these models are released, companies and regulators have limited control over their impact. Therefore, government regulation must address the relative risks and benefits posed by open models.

The title of this report – *Regulating Under Uncertainty: Governance Options for Generative AI* – seeks to convey the unprecedented position of governments as they confront the regulatory challenges AI poses. Regulation is both urgent and unpredictable. It also may be counterproductive, if not done well. However, governments cannot wait until they have perfect and complete information before they act, because doing so may be too late to ensure that the trajectory of technological development does not lead to existential or unacceptable risks. The goal of this report has been to present all of the options that are "on the table" now with the hope that all stakeholders can begin to establish best practices through aggressive information sharing. The risks and benefits of AI will be felt across the entire world. It is critical that the different proposals emerging are assembled in one place so that policy proponents can learn from one another and move ahead in a cooperative fashion.

APPENDICES

Appendix I: Andreessen Horowitz' List of the most popular generative AI applications used by consumers

Top 10 most-used generative AI web products in unique monthly visits		Top 10 most-used generative AI mobile applications in monthly active users	
1-ChatGPT	Chatbot	1-ChatGPT	Chatbot
2-Gemini	Chatbot	2-Microsoft Edge	AI-powered Browser
3-Character.ai	Chatbot imitating historical/fictional characters	3-Photomath	Math camera solver app
4-Liner	End-to-end tool for training machine learning models without code	4-Bing	AI-powered Browser
5-QuillBot	AI-powered writing tool	5-Remini	AI Photo enhancer
6-Poe	Chatbot and multi-model access	6-Brainly	AI Tutor for Homeworks
7-Perplexity	Searchbot	7-Nova	Chatbot
8-JanitorAI	Companion App	8-Chat & Ask AI	Chatbot
9-Civitai	AI-generated content host	9-Facemoji	Original Emoji/Stickers generator
10-Claude	Chatbot	10-Epik	AI Photo enhancer

Source: Andreessen Horowitz: The top 100 generative AI applications

Appendix II: Provisions applicable to general purpose AI models in the European Council's draft of the AI Act[2338]

General Purpose AI Systems (GPAIS)	Common Position on AI Act by Council of the European Union (December 6, 2022)
General principle **Recital 12(c)**	Due to their particular nature and in order to ensure a fair sharing of responsibilities along the AI value chain, GPAIS should be subject to proportionate and more specific requirements and obligations
Definition **Article 3(1)(b)**	● "general purpose AI systems" (GPAIS) means an AI system that, irrespective of how it is placed on the market or put into service, including as open source software, is intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others;" ● "a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems"
Requirements **(Article 4b)**	● Requirements will be specified and adapted in implementing acts in the light of GPAIS characteristics, technical feasibility, specificities of the AI value chain and of market and technological developments ● These requirements include: ○ providing their name and trademark (Article 16 aa); ○ conducting a conformity assessment (Article 16 e); ○ registration (Article 16f); ○ corrective actions (Article 16 g); ○ CE marking (Article 16i); ○ demonstrate conformity (Article 16 j)); ○ appointing an authorized representative (Article 25); ○ EU declaration of conformity (Article 48); ○ post market monitoring (Article 61); ○ sharing information with incoming competitors (Article 4b(5)).
Scope of the requirements **(Article 4a and 4b)**	Requirements apply to GPAIS: ● "which may be used as high risk AI systems or as components of high risk AI system" ● "irrespective of whether the GPAIS is placed on the market or put into service as a pre-trained model and whether further fine-tuning of

[2338] Council of the European Union, Proposal for a regulation laying down harmonized rules on artificial intelligence (**Artificial Intelligence Act)** - General Approach (Dec. 6, 2022) https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf.

	the model is to be performed by the user of the general purpose AI system.”
Exceptions (Article 4c)	● No requirements when the GPAIS provider has explicitly excluded all high-risk uses in the instructions of use or information accompanying the system. ○ Such exclusion shall be made in good faith and shall not be deemed justified if the provider has sufficient reasons to consider that the system may be misused. ● When the GPAIS provider detects or is informed about market misuse they shall take all necessary and proportionate measures to prevent such further misuse, in particular taking into account the scale of the misuse and the seriousness of the associated risks.

Appendix III: Provisions applicable to foundation models in the EU Parliament's draft[2339]

Provisions applicable to Foundation Models	Negotiating Position voted by EU Parliament (June 14, 2023)
Definition of GPAI system -Article 3(1)(d)	an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed.
Definition of foundation model Article 3(1)(c)	an AI system model that is: ● trained on broad data at scale ● is designed for generality of output ● and can be adapted to a wide range of distinctive tasks
Key features of foundation models **Recital (60e and g)**	● AI models developed from algorithms designed to optimize for generality and versatility of output. ● often trained on a broad range of data sources and large amounts of data to accomplish a wide range of downstream tasks, including some for which they were not specifically developed and trained. ● unimodal or multimodal ● trained through various methods, such as supervised learning or reinforced learning ● with specific intended purpose or general purpose AI systems ● re-used in countless downstream AI or general purpose AI systems ● not necessarily high risk AI models ● pre-trained models developed for a narrower, less general, more limited set of applications that cannot be adapted for a wide range of tasks, such as simple multi-purpose AI systems are note foundation models

[2339] Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation on laying down harmonized rules on artificial intelligence (**Artificial Intelligence Act**), https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

Requirements to providers of foundations models **Article 28b**	Providers shall ensure, prior to making the model available on the market or putting it into service, that it is compliant with various requirements: ● **the identification, reduction, and mitigation of reasonably foreseeable risks** to health, safety, fundamental rights, the environment, and democracy and the rule of law, prior to and throughout development ● **quality of data**; appropriate data governance measures, in particular measures to examine the suitability of the data sources and possible biases and appropriate mitigation ● **appropriate levels of performance, predictability, interpretability, corrigibility, safety, and cybersecurity** throughout the lifecycle of the model ● **reduce energy use, resource use and waste,** and increase energy efficiency, and the overall efficiency of the system ● **extensive technical documentation** and intelligible instructions for use ● **quality** management system ● **registration** of the foundation model in the EU database
Precisions about foundation models provided as a service **Recital 60 (f)**	When foundation models are provided as a service, such as through API access: ● the cooperation with downstream providers should extend throughout the time during which that service is provided and supported ● unless ○ the provider of the foundation model transfers the training model as well as extensive and appropriate information on the datasets and the development process of the system ○ the provider restricts the service, such as the API access, in such a way that the downstream provider is able to fully comply with the Regulation without further support from the original provider of the foundation model.

Provisions applicable to systems intended to generate content ("ChatGPT rule") (subset of Foundation Models)	**Negotiating Position voted by EU Parliament (June 14, 2023)**
Definition of systems intended to generate content **Article 28(b)(4)**	foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video ("generative AI")

Obligations of providers of systems intended to generate content	**Article 28(b)(4)** Providers of generative AI models and providers who specialize a foundation model into a generative AI system must: - design and develop systems intended to interact with natural persons in such a way that **the natural person exposed to an AI system is informed that they are interacting with an AI system** in a timely, clear and intelligible manner, unless this is obvious from the circumstances and the context of use. - train, design, and develop the model in such a way as to ensure adequate safeguards against **the generation of content** that is in breach of EUlaw - document and make publicly available a s**ufficiently detailed summary of the use of training data protected under copyright law**.
Obligations of users of systems intended to generate content **Article 52(3)**	**Labeling of machine-generated content** Users of an AI system that generates or manipulates text, audio or visual content ● that would falsely appear to be authentic or truthful ● and which features depictions of people appearing to say or do things they did not say or do, without their consent (‘deep fake’) ● shall disclose in an appropriate, timely, clear and visible manner that the content has been artificially generated or manipulated, as well as, whenever possible, the name of the natural or legal person that generated or manipulated it.

Appendix IV: Proposal of the Spanish Presidency during the Trilogues[2340]

[2340] Luca Bertuzzi, *AI Act: EU countries headed to tiered approach on foundation models amid broader compromise*, EURACTIV (October 17, 2023) https://www.euractiv.com/section/artificial-intelligence/news/spanish-presidency-pitches-obligations-for-foundation-models-in-eus-ai-law/;

Spanish Presidency Proposal	Foundation Models	Very Capable Foundation Models	General Purpose AI systems built on foundation models and used at scale in the EU
Definition	AI model that is capable to competently perform a wide range of distinctive tasks	Foundation models whose capabilities go beyond the current state-of-the-art and may not yet be fully understood. - A foundation model is presumed to be 'very capable' when a given threshold of FLOPs is reached but the presumption is rebuttable -Other possible criteria: * amount of data consumed in training, * potential impact of these foundational models on users, established by the amount of high-risk AI applications that are built on the basis of such foundation models. -AI Office can exceptionally consider a model as "very capable", notably following an investigation.	Thresholds for reach and impact, for example -[10,000] registered business users (i.e. developers) -or [45 million] registered end users
Providers' obligations	1. Before the model is placed on the market: a.document the model and training process, including the results of internal red teaming b."carry out and document model evaluation in accordance with standardized protocols and tools" (i.e. benchmarks). 2. After the model is placed on the market: a. provide information and documentation to the downstream provider b. enable the testing of foundation models by downstream providers. 3.must collaborate with authorities (e.g. the AI Office), who may, upon alert, request the disclosure of the documentation 4.must demonstrate that the models are trained in compliance with applicable EU copyright law, in particular respect the opt-out from the TDM exception. 5. must make publicly available a sufficiently detailed summary about the content used for training and	1.Before the model is placed on the market a. regular external red teaming through vetted red testers b. introducing a risk assessment and mitigation system, also covering possible systemic risks. Code of conduct should be established. 2. After the model is placed on the market: regular compliance controls organized by the AI Office and carried out through independent auditors/researchers.	- regular external red teaming through vetted red testers. - risk assessment and mitigation system, also covering possible systemic risks.

	information about their policies to manage copyright-related aspects.		

Appendix V: Comparative overview of the various Chinese legal frameworks governing generative AI services

Measures	Deep Synthesis (2022)	Interim Measures (2023)	Basic Requirements (2024)
Scope	The application of deep synthesis technology to provide internet information services. (Article 2)	The use of generative AI technologies to provide services to the public in the [mainland] PRC for the generation of text, images, audio, video, or other content. (Article 2)	Generative AI service providers, service providers carrying out safety assessments and improving safety levels, and relevant oversight authorities. (Article 1)
Copyright-ed and personal data used for training	Deep Synthesis providers and technical supporters: • must ensure compliance with applicable regulations regarding the protection of personal information when using such information in training datasets (Article 14) • must prompt the users of the deep synthesis service –when offering editing functions for biometric data, such as faces and voices-- to notify the individuals whose personal information is being edited and obtain their explicit consent. (Article 14)	Generative AI service providers must use data and foundation models that • have lawful sources (Article 7.1) • refrain from infringing intellectual property rights (Article 4.3) • do not infringe on privacy and personal information (Article 4.4) • obtain consent for personal information and comply with relevant laws (Article 7.3)	Generative AI service providers must: • establish an intellectual property rights management strategy and identify any significant intellectual property infringement risks before training (Article 5.2(b)) • obtain the data subject's consent for using personal or sensitive data for training (Article 5.3(c)) or for using user prompts for training (Article 5.1(a)(4)) • prominently disclose the personal information collected and its intended uses (Article 7(b)(2)) • ensure traceability of training data (Article 5.1(c)) • Conduct quality assessments of the training data before training (Article 5.1) • filter out illegal and unhealthy content from the training corpus using methods such as keywords, classification models, and manual spot checks. (Article 5.2)
Data Curation	Deep Synthesis providers and technical supporters are required to	Generative AI service providers must: • increase the quality of training data (Article 7.4)	Generative AI service providers must: - ensure traceability of training data (Article 5.1(c))

	enhance the management and security of training data (Article 14)	• take measures to prevent discrimination during the selection of training data (Article 4.2) • take effective measures to increase the quality, accuracy, precision, objectivity, and diversity of the dataset (Article 7.4) • comply with regulations on cybersecurity, data security, and the protection of personal information (Article 21) • support the establishment of data resources (Article 6) • promote the establishment of public training data resource platforms and the orderly opening of public data by type and grade, expanding high-quality public training data resources (Article 6) If manual tagging is conducted while researching and developing generative AI technology, providers must: - develop and implement clear, specific, and practical tagging rules - undertake a quality assessment of data tagging - conduct sample verification to evaluate the accuracy - conduct necessary training for tagging personnel (Article 8)	- conduct quality assessments of the training data before training (Article 5.1 (a)) - filter out illegal and unhealthy content from the training corpus using methods such as keywords, classification models, and manual spot checks. (Article 5.2)
Technical safeguards and security assessment	Deep Synthesis service providers and technical supporters must: - implement safe and controllable **technical safeguards** (Article 7) - conduct **security assessments (**Article 7) - **regularly audit,** evaluate, and verify the mechanisms of algorithms that produce synthesis (Article 15)	Providers of generative AI services with public opinion properties or the capacity for social mobilization must undergo security assessments in accordance with relevant state provisions. (Article 17)	

	Providers of deep synthesis services **capable of influencing public opinion or mobilizing the public** must conduct a **security assessment** before launching any new products, applications, or features that could impact public opinion or mobilize the public. (Article 20)		
Registry filing	Providers of services capable of influencing public opinion or mobilizing the public must register with relevant regulators (Article 19)	Providers of generative AI services capable of influencing public opinion or mobilizing the public must perform formalities for the filing, modification, or canceling of filings on algorithms. (Article 17)	
User protection	Deep synthesis providers must: - assume primary responsibility for information security and establish systems for personal information protection and prevention of telecommunication network fraud. (Article 7) - prompt users to notify and obtain independent consent from individuals whose biometric information, such as faces and voices, is being edited (Article 14)	Generative AI service providers must: - respect the confidentiality of users' information inputs (Article 11) - clarify the rights and obligations of providers and users through service agreements with users (Article 9) - refrain from collecting unnecessary personal information from users, illegally retaining users' information inputs from which a user's identity can be determined, or illegally providing users' information inputs to third parties (Article 11) - promptly address individuals' requests to access, copy, correct, supplement, or delete personal information (Article 11) - establish a mechanism for receiving and handling complaints from users (Article 15) - guide users' scientific understanding and lawful use of AI and take effective measures to prevent minors from becoming overly reliant on or addicted to generative AI services (Article 10)	Generative AI services providers must: - obtain the data subject's consent for using user prompts for training (Article 5.1(a)(4)) - Implement measures to protect minors: * If the service is not suitable for minors, then technical or administrative measures should be taken to prevent minors' access to the service. (Article 7(a)) * If the service is suitable for minors, guardians should be allowed to implement "anti-addiction measures"(Article 7(a))

Illegal content and activities	Service providers and technical supporters: - are prohibited from using deep synthesis services for creating, reproducing, publishing, or disseminating illegal information or engaging in illegal activities (Article 6) - must promptly identify illegal and harmful information and take effective measures to address such content, - must store related records and report to relevant authorities. (Article 10 and 11) - must establish and complete mechanisms for dispelling rumors and employ measures to dispel the rumors (Article 11)	Providers of generative AI services: - are prohibited from generating fake and harmful information. (Article 4) - bear responsibilities as "content producers" of the AI generated content (Article 9) - should ensure that their services do not generate certain categories of content (Article 4) - must take steps to prevent any illegal activities, take prompt measures to cease the generation and dissemination of illegal content upon discovery, and prevent the recurrence of illegal content (Article 14) - must report incidents to regulators (Article 14) - must rectify the issue for the future, for example by "optimizing training" of models to correct illegal content (Article 14)	Providers of generative AI services must: - Verify and assess the "integrity" of the generated content with regards to the 31 identified risks (Article 8(2)(b)) - must ensure that their services are designed to reject queries that could lead to the generation of prohibited content (Article 7(g)(2))
User monitoring	Deep Synthesis providers : - must develop and publicize rules for technical supporters and users (Article 8) - must establish a user management system, including verifying the real identity of users. (Article 9) - are prohibited from offering services to users who have not undergone identity verification (Article 9)	Generative AI service providers must: - guide users to use generative AI legally and rationally (Article 10) - take steps to prevent any illegal activities by users, including through technical measures such as warnings, by limiting functions available to the user, and by suspending user access to the service (Article 14)	Generative AI providers must: - monitor user prompts with methods, such as keywords and classification models (Article 7(i)) - sanction users that input "illegal or unhealthy" prompts repetitively (Article 7)

Tagging of AI generated work (water-marking)	Deep synthesis providers must: - conspicuously label (watermark) the generated or edited content when such content could confuse the public (Article 17) - include features that allow users to prominently label and alert others regarding their use of such services (Article 17)	Providers of generative AI services must: - add tags on images, videos, and other AI generated content when such content could confuse the public (Article 12)	

Appendix VI: Examples of technical definitions included in the Executive Order 14110

Section 3	Definitions (as provided in the Executive Order)
(d) "AI red-teaming"	"a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Artificial Intelligence red-teaming is most often performed by dedicated 'red teams' that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system."
(j) "differential-privacy guarantee"	"protections that let information about a group to be shared while provably limiting the improper access, use, or disclosure of personal information about particular entities."
(k) "dual-use foundation model"	"an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:" (i) "substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;" (ii) "enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks;" or (iii) "permitting the evasion of human control or oversight through means of deception or obfuscation." "Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities."
(p) "generative AI"	"the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content."
(u) "model weight"	"a numerical parameter within an AI model that helps determine the model's outputs in response to inputs."
(gg) "watermarking"	"the act of embedding information, which is typically difficult to remove, into outputs created by AI —including into outputs such as photos, videos, audio clips, or text — for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance."

Appendix VII: Key studies mandated by Executive Order 14110[2341]

Provision	Implementing agency/department (primary)	Entities covered	Scope or threshold/trigger	Contents
4.5	Commerce	N/A – internal government report and nonbinding guidance	Fostering "capabilities for identifying and labeling synthetic content"	Commerce instructed to: - submit a report "identifying existing and potential future standards, tools, methods, and practices for authenticating, labeling, and detecting synthetic content - as well as techniques for "preventing generative AI from producing" CSAM or "non-consensual intimate imagery of real individuals" - develop guidance (to be periodically updated) on "existing tools and practices" for dealing with synthetic content, including for the purposes listed above - issue guidance (to be periodically updated) for labeling and authenticating content produced by federal government entities
4.6	Commerce	N/A – internal government report	Risks and benefits of "dual-use foundation models (DUFMs) with widely available weights" (basically, powerful open-source models)	Commerce instructed to: - undertake a public consultation to solicit input from stakeholders "on potential risks, benefits, other implications," and to examine "appropriate policy and regulatory approaches" for DUFMs with widely available weights - submit a report on its findings on these topics, including policy and regulatory recommendations

[2341] For further details, see the main text of this report, as well as the White House's fact sheet and the executive order (EO) itself.

4.4(a)	Homeland Security	N/A – internal government report	Potential for AI misuse to enable the development or production of chemical, biological, radiological, and nuclear (CBRN) threats (with "particular focus on biological weapons") - also potential for AI to be used to *counter* CBRN threats	DHS instructed to: - consult with stakeholders to evaluate AI models' capability to pose CBRN threats, as well as options for minimizing risk from such threats - submit a report describing DHS efforts, including: - assessment of the types of AI models that may pose CBRN risks to the U.S. - recommendations for regulating or overseeing the training, deployment, publication, or use of such models, including potential safety evaluations and guardrails that could mitigate threats to national security

Appendix VIII: Instructions to existing agencies in Executive Order 14110[2342]

Provision	Implementing department/agency (primary)	Covered entities	Scope or threshold/trigger	Contents
Suggestions for regulatory agencies to act under existing authority				
8(a)	All independent regulatory agencies	N/A – no immediate or certain impact, but actions ultimately taken by regulatory agencies would cover the private sector	Protecting consumers from fraud, discrimination, privacy violations, and other potential harms that may arise from the use of AI	Agencies encouraged to: - use their respective existing authorities, including administrative rulemaking, to address potential consumer harms from AI - emphasize or clarify where existing regulations and guidance apply to AI - including requirements and expectations related to the transparency of AI models and regulated entities' ability to explain their use of AI models
5.3(a)	FTC	N/A – no immediate or certain impact, but actions ultimately taken by regulatory agencies would cover the private sector	Promoting "competition in AI and related technologies."Sources of competition risk identified in the EO include: - concentrated control of key inputs - unlawful collusion - dominant firms disadvantaging competitors	FTC encouraged to use its existing legal authority, including administrative rulemaking powers, to: - "ensure fair competition in the AI marketplace" - "ensure that consumers and workers are protected" from AI-related harms
4.1(a)	NIST	N/A – non-binding guidance but aimed at industry	Establishing "guidelines and best practices, with the aim of promoting consensus industry	NIST instructed to: - develop a generative AI "companion resource" to its existing "AI Risk Management Framework"

[2342] For further details, see the main text of this report, as well as the White House's fact sheet and the executive order (EO) itself.

			standards, for developing and deploying safe, secure, and trustworthy AI systems”	- develop a generative AI/DUFM companion resource to its existing “Secure Software Development Framework” - launch an “initiative to create guidance and benchmarks for evaluating and auditing AI capabilities” (particularly those that could cause harm) - establish red-teaming guidelines for AI, especially DUFMs

Appendix IX: Examples of purposes for which AI is presumed to be Safety-Impacting and Rights-Impacting (extracts from the OMB Memorandum)[2343]

Purposes That Are Presumed to Have Impact on Safety	**Purposes That Are Presumed to Have Impact on Rights**
A use of AI is presumed to impact safety if, in practical conditions, it controls or significantly affects the outcomes of various agency activities or decisions:	A use of AI is presumed to be have impact on rights if it is used, or is expected to be used, in real-world conditions to control or significantly influence the outcomes of any of the following agency activities or decisions:
-Controlling safety-critical functions within dams, emergency services, electrical grids, energy generation or movement, fire safety systems, food safety mechanisms, traffic control systems and other physical transit systems, water and wastewater systems, or nuclear reactors, materials, and waste. - Maintaining the integrity of elections and voting infrastructure. - Controlling the physical movements of robots or robotic appendages within workplaces, schools, housing, transportation, medical, or law enforcement settings. - Applying kinetic force; delivering biological or chemical agents; or delivering potentially damaging electromagnetic impulses. - Autonomously or semi-autonomously moving vehicles, whether on land, underground, at sea, in the air, or in space. - Controlling the transport, safety, design, or development of hazardous chemicals or biological agents. - Controlling industrial emissions and environmental impacts. - Transporting or managing industrial waste or other controlled pollutants.	- Blocking, removing, hiding, or limiting the reach of protected speech. - In law enforcement contexts, conducting activities such as: - Producing risk assessments about individuals. - Predicting criminal recidivism, criminal offenders, or victims of crime. - Identifying criminal suspects or predicting perpetrators' identities. - Forecasting crime or detecting gunshots. - Tracking personal vehicles over time in public spaces, including using license plate readers. - Conducting biometric identification (e.g., iris, facial, fingerprint, gait matching). - Sketching or reconstructing faces from genetic information. - Monitoring social media or prisons. - Analyzing criminal evidence forensically. - Conducting forensic genetics, cyber intrusions, or physical location monitoring. - Making determinations related to sentencing, parole, bail, or detention. - Deciding or providing risk assessments related to immigration, asylum, or detention status, including: - Assessing risks for individuals traveling to or within

[2343] Shalanda D. Young, *Advancing Governance Innovation and Risk Management for Agency Use of Artificial Intelligence, Memorandum for the Heads of Executive Departments and Agencies*, Office of Management and Budget (OMB) (March 28, 2024) Appendix I, https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf

- Designing, constructing, or testing industrial equipment, systems, or structures that could pose a significant safety risk if they failed. - Carrying out medically relevant functions of medical devices; providing medical diagnoses; determining medical treatments; conducting health-risk assessments, including for medical insurance, drug addiction, suicide risk, or other violence; detecting or preventing mental health issues; flagging patients for interventions; allocating care in public insurance contexts; or managing health-insurance costs and underwriting. - Detecting the presence of dangerous weapons or violent acts. - Deciding when to summon first responders to an emergency. - Controlling access to or security of government facilities. - Determining or implementing enforcement actions related to sanctions, trade restrictions, or other controls on exports, investments, or shipping.	the U.S. - Monitoring physical locations for immigration and detention purposes. - Forecasting individual migration activity. - Conducting biometric identification for one-to-many identification in publicly accessible spaces. - Detecting or measuring emotions, thought, impairment, or deception in humans. - Replicating a person's likeness or voice without express consent. - In educational contexts, conducting activities such as: - Detecting student cheating or plagiarism. - Influencing admissions processes or monitoring students online. - Projecting student progress, recommending disciplinary interventions. - Determining access to educational resources or eligibility for student aid. - Screening or monitoring tenants in public housing; providing home valuations; underwriting mortgages; determining home insurance terms. - Determining terms or conditions of employment, including: - Conducting pre-employment screening, reasonable accommodation, or performance management. - Making decisions on hiring, pay, promotion, or termination. - Recommending disciplinary action or conducting workplace surveillance. - Carrying out medically relevant functions such as: - Providing medical diagnoses or determining medical treatments. - Conducting health-risk assessments for medical or insurance purposes. - Detecting or preventing mental-health issues or allocating care in public insurance. - Allocating loans; determining access to financial systems; credit scoring; making financial audit decisions; assessing insurance risks; determining interest rates or financial penalties. - Making decisions regarding access to, eligibility for, or revocation of critical government resources or services, including: - Allowing or denying access through biometrics to IT systems. - Detecting fraudulent use of government services.

	- Assigning penalties in the context of government benefits. - Translating between languages for official communications where responses are legally binding; providing live language interpretation or translation without a competent interpreter for legally significant interactions. - Providing recommendations, decisions, or risk assessments about: - Adoption matching, child protective actions, child custody determinations. - Protective actions for senior citizens or disabled persons.

Appendix X: The 11 recommendations of the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems

Principles	Content
1-Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate, and mitigate risks across the AI lifecycle.	● Developers and users commit to ongoing assessments of risk and reliability in their own systems. ● Red teaming is recommended. ● Several risk categories are identified as requiring special attention: o Chemical, biological, and nuclear weapons development and use o Cyberattacks o Human rights and democratic values (stopping disinformation and protecting privacy) o Health and health infrastructures o Avoiding chain reactions of negative events o Protect against autonomous system and loss of human control
2-Identify and mitigate vulnerabilities and, where appropriate, incidents and patterns of misuse, after deployment, including placement on the market.	Urge organizations to monitor vulnerabilities and stay ahead of issues after deployment. Text suggests bounty systems as a way to incentivize third party users to help identify system gaps. The main idea is to hold organizations accountable for the post-market effects of their models and encourage documentation of all reported incidents.
3-Publicly report advanced AI systems' capabilities, limitations, and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increased accountability	Calls for producing public, up-to-date transparency reports to identify risks and highlight tested solutions in developed models.
4-Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia	Encourages companies to share evaluation reports and information on security and safety risks in a way that respects intellectual property rights but keeps the public informed. Public authorities should also have information needed for comprehensive risk management. Organizations should work together to develop shared standards and mechanisms.

5-Develop, implement, and disclose AI governance and risk management policies, grounded in a risk-based approach – including privacy policies and mitigation measures	Organizations must establish and disclose their AI governance policies and organizational mechanisms for managing risk. This includes, especially, implementing and disclosing privacy policies.
6-Invest in and implement robust security controls, including physical security, cybersecurity, and insider threat safeguards across the AI lifecycle.	Organizations should establish robust internal risk detection programs, guaranteeing control and protection of the weights of models and algorithms, servers, and datasets to prevent cyberattacks.
7-Develop and deploy, where technically feasible, reliable content authentication and provenance mechanisms, such as watermarking or other techniques, to enable users to identify AI-generated content.	Content authentication, user authentication, watermarking, etc.
8-Prioritize research to mitigate societal, safety, and security risks and prioritize investment in effective mitigation measures.	Share decisive results in this area and prioritize investments in safety.
9-Prioritize the development of advanced AI systems to address the world's greatest challenges, notably —but not limited to— the climate crisis, global health, and education.	Ensure AI systems remain in line with the United Nations' sustainable development goals.
10-Advance the development of and, where appropriate, adoption of international technical standards.	Organizations should work with standardization bodies (Standards Development Organizations) and coordinate among themselves to develop both common, readable standards and common tools that can be used by the general public to identify AI-generated content.
11-Implement appropriate data input measures and protections for personal data and intellectual property	Encourage any and all efforts to erase harmful bias.