PITFALLS OF EVIDENCE-BASED AI POLICY

Stephen Casper
MIT CSAIL
scasper@mit.edu

David Krueger
Mila

Dylan Hadfield-Menell
MIT CSAIL

arXiv:2502.09618v1 [cs.CY] 13 Feb 2025

At this very moment, I say we sit tight and assess.
– President Janie Orlean, Don't Look Up

ABSTRACT

Nations across the world are working to govern AI. However, from a technical perspective, there is uncertainty and disagreement on the best way to do this. Meanwhile, recent debates over AI regulation have led to calls for "evidence-based AI policy" which emphasize holding regulatory action to a high evidentiary standard. Evidence is of irreplaceable value to policymaking. However, holding regulatory action to too high an evidentiary standard can lead to systematic neglect of certain risks. In historical policy debates (e.g., over tobacco ca. 1965 and fossil fuels ca. 1985) "evidence-based policy" rhetoric is also a well-precedented strategy to downplay the urgency of action, delay regulation, and protect industry interests. Here, we argue that if the goal is evidence-based AI policy, the first regulatory objective must be to actively facilitate the process of identifying, studying, and deliberating about AI risks. We discuss a set of 15 regulatory goals to facilitate this and show that Brazil, Canada, China, the EU, South Korea, the UK, and the USA all have substantial opportunities to adopt further evidence-seeking policies.

CONTENTS

1 How do We Regulate Emerging Tech? 2

1.1 "Nope, I'm against evidence-based policy." . . . 2
1.2 A Broad, Emerging Coalition . . . 2
1.3 A Vague Agenda? . . . 3

2 The Evidence is Biased 4

2.1 Selective Disclosure . . . 4
2.2 Easy- vs. Hard-to-Measure Impacts . . . 4
2.3 Precedented vs. Unprecedented Impacts . . . 5
2.4 Ingroups vs. Outgroups . . . 5
2.5 The Culture & Values of the AI Research Community . . . 5
2.6 Industry's Entanglement with Research . . . 6

3 Lacking Evidence as a Reason to Act 8

3.1 Substantive vs. Process Regulation . . . 9
3.2 In Defense of Compute & Cost Thresholds in AI Regulation . . . 9

4 We Can Pass Evidence-Seeking Policies Now **9**

4.1 15 Evidence-Seeking AI Policy Objectives 10

4.2 Ample Room for Progress 10

4.3 The Duty to Due Diligence from Discoverable Documentation of Dangerous Deeds 12

4.4 Considering Counterarguments 12

5 Building a Healthier Ecosystem **13**

1 How do We Regulate Emerging Tech?

Recently, debates over AI governance have been ongoing across the world. A common underlying theme is the challenge of regulating emerging technologies amidst uncertainty about the future. Even among people who strongly agree that it is important to regulate AI, there is sometimes disagreement about when and how. This uncertainty has led some researchers to call for "evidence-based AI policy."

1.1 "Nope, I'm against evidence-based policy."

See how awful that sounds? This highlights a troublesome aspect of how things are sometimes framed. Of course, evidence is indispensable. But there is a pitfall of holding policy action to too high an evidentiary standard:

> **Postponing regulation that enables more transparency and accountability on grounds that it's "not evidence-based" is counterproductive.**

As we will argue, focusing too much on getting evidence before we act can paradoxically make it harder to gather the information we need.

1.2 A Broad, Emerging Coalition

Recently, there have been a number of prominent calls for evidence-based AI policy. For example, several California congressmembers and Governor Gavin Newsom recently argued against an AI regulatory bill in California by highlighting that it was motivated by mitigating future risks that have not been empirically observed:

> *There is little scientific evidence of harm of 'mass casualties or harmful weapons created' from advanced models.*
>
> – Zoe Lofgren et al. in an open letter to Gavin Newsom

> *[Our] approach... must be based on empirical evidence and science... [we need] Al risk management practices that are rooted in science and fact.*
> – Gavin Newsom in his veto of bill SB 1047

Others in academia have echoed similar philosophies of governing AI amidst uncertainty. For example, in their book AI Snake Oil (Narayanan & Kapoor, 2024), Princeton researchers Arvind Narayanan and Sayash Kapoor claim:

> *The whole idea of estimating the probability of AGI risk is not meaningful... We have no past data to calibrate our predictions.*
> – Narayanan & Kapoor (2024), AI Snake Oil

They follow this with an argument against the precautionary principle (Taleb et al., 2014), claiming that policymakers should take a noncommittal approach in the face of uncertainty and not act on speculative estimates of future AI risks.

Meanwhile, Jacob Helberg, a senior adviser at the Stanford University Center on Geopolitics and Technology, has argued that there just isn't enough evidence of AI discrimination to warrant policy action.

> *This is a solution in search of a problem that really doesn't exist... There really hasn't been massive evidence of issues in AI discrimination.*
> – Jacob Helberg on priorities for the current presidential administration

And the seventeen authors of a recent article titled, *A Path for Science- and Evidence-Based AI Policy*, argue that:

> *AI policy should be informed by scientific understanding... if policymakers pursue highly committal policy, the... risks should meet a high evidentiary standard.*
> – Bommasani et al. (2024a), A Path for Science- and Evidence-based AI Policy

Overall, the evidence-based AI policy coalition is diverse. It includes a variety of policymakers and researchers who do not always agree with each other. We caution against developing a one-dimensional view of this coalition or jumping to conclusions from quotes out of context. However, **this camp is generally characterized by a desire to avoid pursuing highly committal policy absent compelling evidence.**

1.3 A Vague Agenda?

Calls for evidence-based policy are not always accompanied by substantive recommendations. However, Bommasani et al. (2024a) end their article with a set of four milestones for researchers and policymakers to pursue:[1]

> **Milestone 1:** A taxonomy of risk vectors to ensure important risks are well-represented
> **Milestone 2:** Research on the marginal risk of AI for each risk vector
> **Milestone 3:** A taxonomy of policy interventions to ensure attractive solutions are not missed
> **Milestone 4:** A blueprint that recommends candidate policy responses to different societal conditions

These milestones are extremely easy to agree with. Unfortunately, they are also unworkably vague. It is unclear what it would mean for them to be accomplished. In fact, for these milestones, it is not hard to argue that existing reports reasonably meet them. For example, the AI Risk Repository (Slattery et al., 2024) predates Bommasani et al. (2024a) and offers a meta-review, taxonomy, and living database of AI risks discussed in the literature. If this does not offer a workable taxonomy of risks (Milestone 1), it is unclear what would.[2]

These milestones are an encouraging call to actively improve our understanding. However, absent more precision, we worry that similar arguments could be misused as a form of tokenism to muddy the waters and stymie policy action. In the rest of this paper, we will argue that holding regulatory action to too high an evidentiary standard can paradoxically make it harder to gather the information that we need for good governance.

[1] Bommasani et al. (2024a) also call for the establishment of a registry, evaluation, red-teaming, incident reporting, and monitoring but do not specify any particular role for regulators to play in these. They also make a nonspecific call for policymakers to broadly invest in risk analysis research and to investigate transparency requirements.

[2] For milestone 2, most relevant research is domain-specific; consider Metta et al. (2024), Sandbrink (2023), Musser (2023), and Cazzaniga et al. (2024) as examples. Note, however, that forecasting future marginal risks will always be speculative to some degree. See also Bengio et al. (2025). Meanwhile, milestones 3 and 4 essentially describe the first and second stages of the AI regulation process, so existing regulatory efforts already are working on these (e.g., Arda, 2024).

2 THE EVIDENCE IS BIASED

In its pure form, science is a neutral process. But it is never done in a vacuum. Beneath the cloak of objectivity, there are subjective human beings working on problems that were not randomly selected (Kuhn, 1962). There is a laundry list of biases subtly shaping the evidence produced by AI researchers. A policymaking approach that fixates on existing evidence to guide decision-making will systematically neglect certain problems.

2.1 SELECTIVE DISCLOSURE

In February 2023, Microsoft announced Bing Chat, an AI-powered web browsing assistant. It was a versatile, semi-autonomous copilot to help users browse the web. It was usually helpful, but sometimes, it went off the rails. Users found that it occasionally took on shockingly angsty, deceptive, and outright aggressive personas. It would go so far as to threaten users chatting with it. Rest assured, everyone was fine. Bing Chat was just a babbling web app that could not directly harm anyone. But it offers a cautionary tale. Right now, developers are racing to create increasingly agentic and advanced AI systems (Chan et al., 2023; Casper et al., 2025). If more powerful future systems go off the rails in similar ways, it could spell trouble.

Following the Bing Chat incidents, Microsoft's public relations strategy focused on patching the issues and moving on. To the dismay of many AI researchers, Microsoft never published a public report on the incident. If Microsoft had nothing but humanity's best interests at heart, it could substantially help researchers by reporting on the technical and institutional choices that led to Bing Chat's behaviors. However, it's just not in their public relations interests to do so.

Historically, AI research and development has been a very open process. For example, code, models, and methodology behind most state-of-the-art AI systems were broadly available pre-2020. More recently, however, developers like Microsoft have been exercising more limited and selective transparency (Bommasani et al., 2024b). **Due to a lack of accountability in the tech industry, some lessons remain simply out of reach. There is a mounting crisis of transparency in AI when it is needed the most.**

2.2 EASY- VS. HARD-TO-MEASURE IMPACTS

The scientific process may be intrinsically neutral, but not all phenomena are equally easy to study. Most of the downstream societal impacts of AI are difficult to accurately predict in a laboratory setting. The resulting knowledge gap biases purely evidence-based approaches to neglect some issues simply because they are difficult to study.

> *Thoroughly assessing downstream societal impacts requires nuanced analysis, interdisciplinarity, and inclusion... there are always differences between the settings in which researchers study AI systems and the ever-changing real-world settings in which they will be deployed.*
> – Bengio et al. (2024), International Scientific Report on the Safety of Advanced AI

Differences in the measurability of different phenomena can cause insidious problems to be neglected. For instance, compare explicit and implicit social biases in modern language models. Explicit biases from LLMs are usually easy to spot. For example, it is relatively easy to train a language model against expressing harmful statements about a demographic group. But even when we do this to language models, they still consistently express more subtle biases in the language and concept associations that they use to characterize different people (Wan et al., 2023; Wan & Chang, 2024; Bai et al., 2024; Hofmann et al., 2024).

Meanwhile, *benchmarks* provide the main scaffolding behind research progress in AI (Patterson, 2012; Hendrycks & Woodside, 2022). For example, tests like GPQA (Rein et al., 2023) actively serve to guide progress on language model capabilities. Many of the benchmarks used in AI research are designed with the hope that they can help us understand downstream societal impacts. However, the strengths of benchmarks are also their weaknesses. Standardized, simplified, and portable measures of system performance often make for poor proxies to study real-world impacts

(Raji et al., 2021). For example, in a systematic study of benchmarks designed to assess harmlessness in AI systems, Ren et al. (2024) found that many existing benchmarks intended to evaluate these qualities were, in practice, more reflective of a model's general capabilities than anything else.

2.3 Precedented vs. Unprecedented Impacts

In the history of safety engineering, many major system failures follow a certain story (Dekker, 2019). It starts off with some system – e.g., a dam, bridge, power plant, oil rig, building, etc – that functions normally for a long time. At first, this is accompanied by direct evidence of benefits and no evidence of major harms which can lull engineers into a false sense of security. But tragedy often strikes suddenly. For example, before the infamous 1986 Challenger space shuttle explosion, there were 9 successful launches (Gebhardt, 2011), which was a factor that led engineers to neglect safety warnings before the infamous 10th launch. Things were fine, and the empirical evidence looked good until disaster struck. AI is a very powerful technology, and if it ever has a 'Chernobyl moment,' a myopic focus on empirical evidence would be the exact way to lead us there.

2.4 Ingroups vs. Outgroups

The AI research community does not represent humanity well. For example, AI research is dominated by White and Asian (AI Index Steering Committee, 2021), English-speaking (Singh et al., 2024) men (Abdulla & Chahal, 2023). It is also relatively culturally homogenous:

> *Since AI technologies are mostly conceived and developed in just a handful of countries, they embed the cultural values and practices of these countries.*
> – Prabhakaran et al. (2022), Cultural Incongruities in Artificial Intelligence

For example, AI ethics researchers have contrasted India and the West to highlight the challenges posed by cultural homogeneity in the research community. In the West, societal discussions around fairness can, of course, be very nuanced, but they reflect Western experiences and are often characterized by a focus on race and gender politics. In India, however, the axes of social disparity are different and, in many ways, more complex. For example, India has 22 official languages, a greater degree of religious conflict, and a historical Caste system. This has led researchers to argue that the AI community is systematically poised to neglect many of the challenges in India and other non-Western parts of the world (Sambasivan et al., 2021; Bhatt et al., 2022; Qadri et al., 2023).

2.5 The Culture & Values of the AI Research Community

Perhaps the most important ingroup/outgroup contrast to consider is the one between the AI research community and the rest of humanity. It is clear that AI researchers do not demographically represent the world (AI Index Steering Committee, 2021; Singh et al., 2024; Abdulla & Chahal, 2023). Meanwhile, they tend to have much more wealth and privilege than the vast majority of the rest of the world. And they tend to be people who benefit from advances in technology instead of being historically or presently marginalized by it. This prompts a serious question:

Is the AI research community prepared to put people over profit and performance?

In their paper, *The Values Encoded in Machine Learning Research*, Birhane et al. (2022) analyzed 100 prominent, influential machine learning papers from 2008 to 2019. They annotated each based on what values were reflected in the paper text. The results, shown in Figure 1, revealed a red flag. They found an overwhelming predominance of values pertaining to system performance (green) over the other categories of user rights and ethical principles. This suggests that the AI community may be systematically predisposed to produce evidence that will disproportionately highlight the benefits of AI compared to its harms.

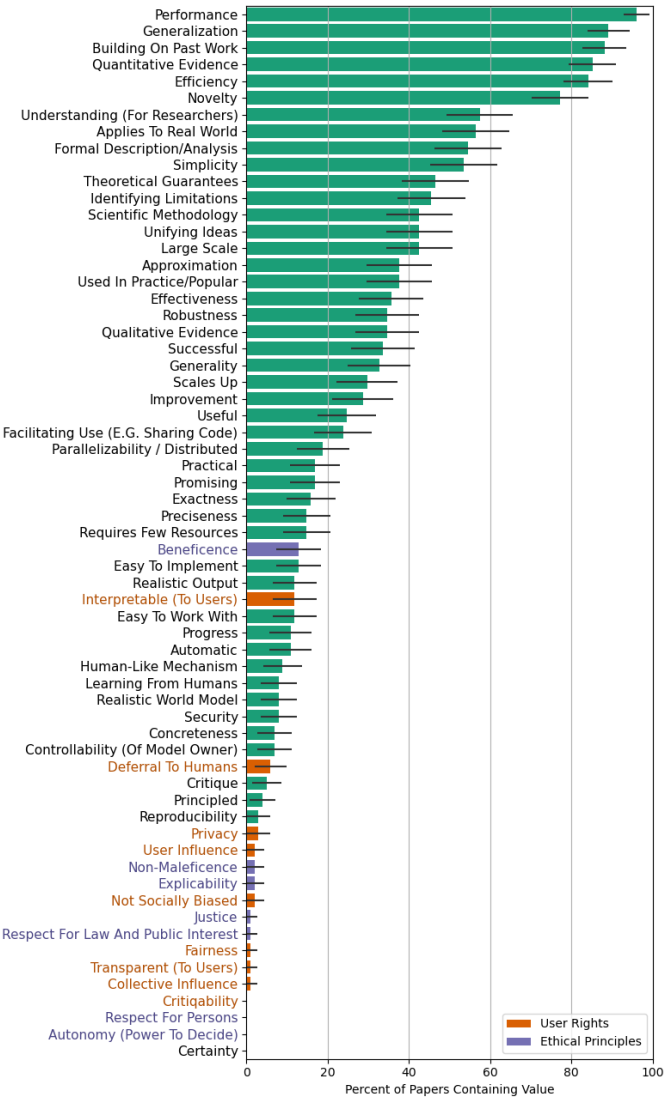


Figure 1: From Birhane et al. (2022), *The Values Encoded in Machine Learning Research*. Among the values represented in prominent AI research papers, there is an overwhelming predominance of ones pertaining to technical system performance. Is the AI research community prepared to put human interests over system performance?

2.6 Industry's Entanglement with Research

Who is doing the AI research? Where is the money coming from? In many cases, the answer to both is the tech companies that would be directly affected by regulation. For instance, consider the NeurIPS conference. Google DeepMind, Microsoft, and Meta all ranked in the top 20 organizations by papers accepted (Figure 2).

The reach of industry labs into the research space involves more than just papers. **AI academia's entanglement with industry runs deep:**

> *Imagine if, in mid-December of 2019, over 10,000 health policy researchers made the yearly pilgrimage to the largest international health policy conference in the world. Among the many topics discussed... was how to best deal with the negative effects of increased tobacco usage... Imagine if many of the speakers who graced*

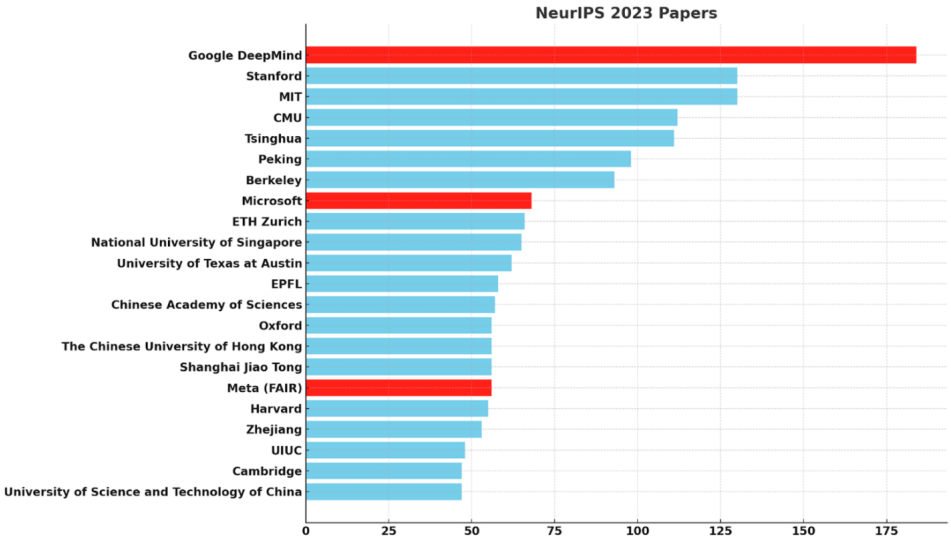


Figure 2: Paper count by organization from the NeurIPS 2023 conference. AI companies directly influence the evidence base.

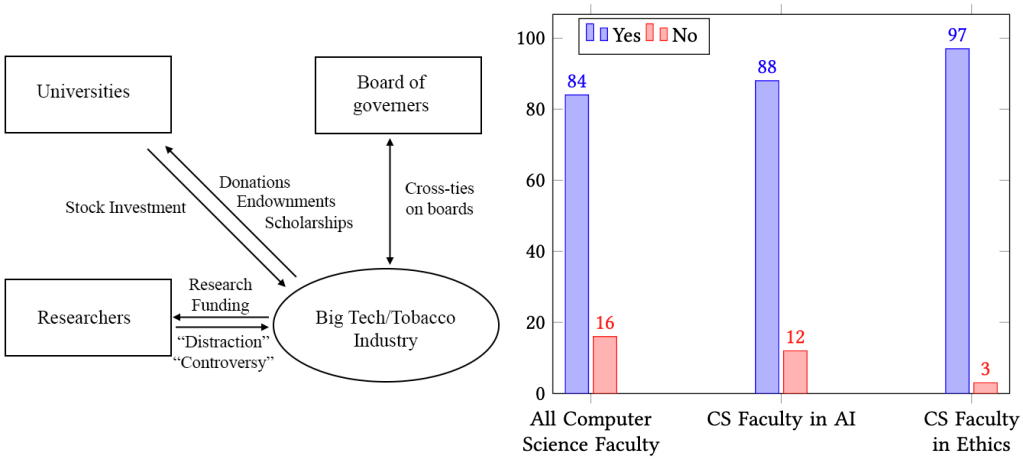


Figure 3: From Abdalla & Abdalla (2020) (Left) There are striking similarities between the anti-regulatory influences of Big Tobacco on public health research and Big Tech on AI research. (Right) The large majority of academic CS faculty have at some point received direct funding/awards from Big Tech or have been employed by Big Tech.

> *the stage were funded by Big Tobacco. Imagine if the conference itself was largely funded by Big Tobacco.*
> – A discussion alluding to the NeurIPS 2019 conference from Abdalla & Abdalla (2020), The Grey Hoodie Project Big Tobacco, Big Tech, and the threat on academic integrity

When a powerful industry is facing regulation, it is in its interest to pollute the evidence base and public discussion around it in order to deny risks and delay action. **A key way that this manifests is with assertions that we need more evidence and consensus before we act.**

> *Is it any wonder that those who benefit the most from continuing to do nothing emphasize the controversy among scientists and the need for continued research?*
> – Giere et al. (2006), Understanding Scientific Reasoning

A common '**Deny and Delay Playbook**' has been used in historical policy debates to delay meaningful regulation until long after it was needed. A common story has played out in debates around tobacco, acid rain, the ozone layer, and climate change (Oreskes & Conway, 2010). In each case, industry interests pushed biased science to cast doubt on risks and made noise in the media about how there just wasn't enough evidence to act yet. This represents a misuse of the scientific process. Of course, all scientific theories are tentative and subject to criticism – this is exactly why science is so useful. But doubtmongering can be abused against the public interest. Calls for evidence-based policy are often used more as a rhetorical pretext for kicking the can down the road than as a constructive, prescriptive proposal for action (Cairney, 2021).

> *Any evidence can be denied by parties sufficiently determined, and you can never prove anything about the future; you just have to wait and see.*
> – Oreskes & Conway (2010), Merchants of Doubt

To illustrate this, we invite the reader to speculate about which of these quotes came from pundits recently discussing AI regulation and which came from merchants of doubt for the tobacco and fossil fuel industries. Who said what? Answers in footnote.[3]

There is no need of going off [without a thorough understanding] and then having to retract... We should take no action unless it can be supported by reasonably positive evidence.	*In addition to its misplaced emphasis on hypothetical risks, we are also concerned that [redacted] could have unintended consequences [on U.S. competitiveness]...It may be the case that the risks posed by [redacted] justify this precaution. But current evidence suggests otherwise.*	*The scientific base for [redacted] includes some facts, lots of uncertainty, and just plain ignorance; it needs more observation... There is also major disagreement... The scientific base for [redacted] is too uncertain to justify drastic action at this time.*

To see an example of Big Tech entangled with calls for "evidence-based AI policy," we need to look no further than Bommasani et al. (2024a): A Path for Science- and Evidence-based AI Policy (discussed above in Section 1). Five of its seventeen authors have undisclosed for-profit industry affiliations.[4] These include the vice president of AI research at Meta and cofounders of World Labs, http://together.ai, Databricks, Anyscale, and :probabl, each of which might be affected by future AI regulations.[5] Failing to disclose clear conflicts of interest in an explicitly political article fails to meet standards for ethical disclosure in research. These standards exist for good reason because a policymaker reading *A Path for Science- and Evidence-based AI Policy* (Bommasani et al., 2024a) might interpret it very differently if it were clear that some of the authors had obvious conflicts of interest. It is certainly a red flag that calls for more evidence before passing highly committal regulation are coming, in part, from authors with conveniently hidden industry ties.

3 Lacking Evidence as a Reason to Act

So if the evidence is systematically biased? What do we do? How do we get more, better evidence?

[3] (Left) A cancer doctor testifying to the US congress in 1965 on tobacco and public health. (Middle) Zoe Lofgren and other representatives in a 2024 open letter to Gavin Newsom on AI regulation. (Right) Fred Singer in a paper arguing against climate action in 1990.

[4] Li, Pineau, Varoquaux, Stoica, Liang

[5] The original version of the article did not contain any disclaimers about omitted author affiliations. However, it was updated in October 2024 to disclaim that "Several authors have unlisted affiliations in addition to their listed university affiliation. This piece solely reflects the authors' personal views and not those of any affiliated organizations." However, these conflicts of interest are still not disclosed.

3.1 SUBSTANTIVE VS. PROCESS REGULATION

We argue that a need to more thoroughly understand AI risks is a reason to pass regulation–not to delay it. To see this, we first need to understand the distinction between "substantive" regulation and "process" regulation. For our purposes, we define them as such:

Substantive regulation limits **what** things developers can do with their AI systems.	**Process** regulation limits **how** developers do what they do with their AI systems.

These two categories of regulations do not only apply to AI. In the food industry, for example, ingredient bans are substantive regulations while requirements for nutrition facts are process regulations. Process regulations usually pose significantly lower burdens and downsides than substantive ones. The key reason why this distinction is important is that, as we will argue:

> **A limited scientific understanding can be a legitimate (but not necessarily decisive) argument to postpone substantive regulation. But the exact opposite applies to process regulation.**

Depending on whether we are considering substantive or process regulation, the argument can go different ways. To see an example, let's consider some recent discussions on cost and compute thresholds in AI regulations.

3.2 IN DEFENSE OF COMPUTE & COST THRESHOLDS IN AI REGULATION

Some AI policy proposals set cost and compute thresholds such that, if a system's development surpasses these, it would be subject to specific requirements. Some researchers have rightly pointed out that there are hazards associated with this; cost and compute can be poor proxies for societal risk (Hooker, 2024; Heim & Koessler, 2024).

These are important and needed points about the limitations of cost and compute thresholds. For example, suppose that we are considering substantive regulations that prevent deploying certain models in certain ways. In this case, we would need careful cost-benefit analysis and the ability to adapt regulatory criteria over time. But until we have government agencies who are capable of performing high-quality evaluations of AI systems' risks, cost and compute thresholds may be the only tenable proxy available (Heim & Koessler, 2024).

In the case of process regulation, there is often a lack of substantial downside. For example, consider policies that require developers to register a system with the government if its development exceeds a cost or compute threshold. Compared to inaction, the upside is a significantly increased ability of the government to monitor the frontier model ecosystem. As for the downside? Sometimes certain companies will accidentally be required to do more paperwork than regulators may have intended. Compared to the laundry list of societal-scale risks from AI (Slattery et al., 2024), we can safely say that this risk is practically negligible.

4 WE CAN PASS EVIDENCE-SEEKING POLICIES NOW

It is important to understand the role of process regulation in helping us to get evidence, especially since governments often tend to underinvest in evidence-seeking during institutional design (Stephenson, 2011). In contrast to vague calls for more research, we argue that a truly evidence-based approach to AI policy is one that proactively helps to produce more information.

> **If we want "evidence-based" AI policy, our first regulatory goal must be producing evidence. We don't need to wait before passing process-based, risk-agnostic AI regulations to get more actionable information.**

4.1 15 EVIDENCE-SEEKING AI POLICY OBJECTIVES

Here, we outline a set of AI regulatory goals related to **institutions**, **documentation**, **accountability**, and **risk-mitigation practices** designed to produce more evidence. Each is process-based and fully risk-agnostic. We argue that the current lack of evidence about AI risks is not a reason to delay these, but rather, a key reason why they are useful.

1. **AI governance institutes**: National governments (or international coalitions) can create AI governance institutes to research risks, evaluate systems, and curate best risk management practices that developers are encouraged to adhere to.
2. **Model registration**: Developers can be required to register (McKernon et al., 2024) frontier systems with governing bodies (regardless of whether they will be externally deployed).
3. **Model specification and basic info**: Developers can be required to document intended use cases and behaviors (e.g., OpenAI, 2024) and basic information about frontier systems such as scale.
4. **Internal risk assessments**: Developers can be required to conduct and report on internal risk assessments of frontier systems.
5. **Independent third-party risk assessments**: Developers can be required to have an independent third-party conduct and produce a report (including access, methods, and findings) on risk assessments of frontier systems (Raji et al., 2022; Anderljung et al., 2023; Casper et al., 2024). They can also be required to document if and what "safe harbor" policies they have to facilitate independent evaluation and red-teaming (Longpre et al., 2024).
6. **Plans to minimize risks to society**: Developers can be required to produce a report on risks (Slattery et al., 2024) posed by their frontier systems and risk mitigation practices that they are taking to reduce them.
7. **Post-deployment monitoring reports**: Developers can be required to establish procedures for monitoring and periodically reporting on the uses and impacts of their frontier systems.
8. **Security measures**: Given the challenges of securing model weights and the hazards of leaks (Nevo et al., 2024), frontier developers can be required to document high-level non-compromising information about their security measures (e.g., Anthropic, 2024).
9. **Compute usage**: Given that computing power is key to frontier AI development (Sastry et al., 2024), frontier developers can be required to document their compute resources including details such as the usage, providers, and the location of compute clusters.
10. **Shutdown procedures**: Developers can be required to document if and which protocols exist to shut down frontier systems that are under their control.
11. **Documentation availability**: All of the above documentation can be made available to the public (redacted) and AI governing authorities (unredacted).
12. **Documentation comparison in court**: To incentivize a race to the top where frontier developers pursue established best safety practices, courts can be given the power to compare documentation for defendants with that of peer developers.
13. **Labeling AI-generated content**: To aid in digital forensics, content produced from AI systems can be labeled with metadata, watermarks, and notices.
14. **Whistleblower protections**: Regulations can explicitly prevent retaliation and offer incentives for whistleblowers who report violations of those regulations.
15. **Incident reporting**: Frontier developers can be required to document and report on substantial incidents in a timely manner.

4.2 AMPLE ROOM FOR PROGRESS

As we write this in February 2025, parallel debates over AI safety governance are unfolding across the world. There are a number of notable existing and proposed policies.

- **Brazil** passed Bill No. 2338 of 2023 (Brazil, 2023) (enacted) on regulating the use of Artificial Intelligence, including algorithm design and technical standards in December 2024.

- **Canada** recently established an AI Safety Institute (exists), and its proposed AI and Data Act (Canada, 2022) (proposed) is currently under consideration in House of Commons Committee.
- **China** has enacted its Provisions on the Administration of Deep Synthesis Internet Information Services (China, 2022b) (enacted), Provisions on the Management of Algorithmic Recommendations in Internet Information Services (China, 2022a) (enacted), and Interim Measures for the Management of Generative AI Services (China, 2023) (enacted). There are also working drafts of a potential future 'The Model Artificial Intelligence Law' (CASS, 2023) (proposed).
- In the **European Union**, the EU AI Act (EU, 2024) (enacted) was passed in March 2024, and a large undertaking to design specific codes of practice is underway.
- **South Korea** passed the [Act on the Development of Artificial Intelligence and Establishment of Trust (AI Basic Act) (National Assembly of the Republic of Korea, 2025) in December 2024.
- The **UK**'s AI Safety Institute (exists) is currently building capacity and partnerships to evaluate risks and establish best risk-management practices. Thus far, the UK's approach to AI regulation has been non-statutory (but new draft legislation may exist within a few months).
- In the **United States**, Donald Trump overturned Executive Order 14110 (Biden, 2023) after assuming office in January 2025. This may or may not lead the US AI Safety Institute (exists) to be shut down. It also might permanently stall a potential policy (Department of Commerce, 2024) on model and compute reporting that the Department of Commerce proposed in response to the executive order. Meanwhile, the AI Advancement and Reliability Act (USA, 2024a) (proposed) was drafted last congress and will be re-introduced this congress. Finally, the Future of AI Innovation Act (USA, 2024b) (drafted) and the Preserving American Dominance in Artificial Intelligence Act (S56, 2024) (drafted) were also introduced last congress. However, as of February 2025, they are currently simply drafts.

So how are each of these countries faring?

	Brazil	Canada	China	EU	Korea	UK	USA
1. AI governance institutes	O	✔	O*	✔	✔	✔	✔
2. Model registration	✗	✗	✔	✔	✔	✗	O*
3. Model specification and basic info	O	O*	O	✔	O	✗	✗
4. Internal risk assessments	✔	O*	O	✔	✔	✗	✗
5. Independent 3rd-party risk assessments	✗	✗	O	O	O	✗	✗
6. Plans to minimize risks to society	O	✗	O	✔	O	✗	✗
7. Post-deployment monitoring reports	✗	✗	✗	✔	✗	✗	✗
8. Security measures	✗	✗	O	✔	✗	✗	✗
9. Compute usage	✗	✗	✗	O	✗	✗	O*
10. Shutdown procedures	✔	O*	✗	O	✗	✗	✗
11. Documentation availability	✗	O*	O	O	O	✗	✗
12. Documentation comparison in court	✗	✗	✗	✗	✗	✗	✗
13. Labeling AI-generated content	✗	✗	✔	O	O	✗	✗
14. Whistleblower protections	✗	✗	✔	✔	✗	✗	✗
15. Incident reporting	✔	✗	✗	✔	✗	✗	✗

Table 1: ✔= Yes, O= Partial, ✗= No, ∗ = proposed but not enacted. There is significant room for progress across the world on passing evidence-seeking AI policy measures. See details on each row above. Note that this table represents a snapshot in time (February 2025). In the USA, we omit the two bills ((USA, 2024b; S56, 2024)) discussed above that were proposed last congress but have not been scheduled for re-introduction this congress.

4.3 THE DUTY TO DUE DILIGENCE FROM DISCOVERABLE DOCUMENTATION OF DANGEROUS DEEDS

The objectives outlined above hinge on documentation. 2-10 are simply requirements for documentation, and 11-12 are accountability mechanisms to ensure that the documentation is not perfunctory. This is no coincidence. When it is connected to external scrutiny, documentation can be a powerful incentive-shaping force. Under a robust regime implementing the above the public and courts could assess the quality of developers' risk management practices. As such, this type of regulatory regime could incentivize a race to the top on risk-mitigation standards (Hadfield & Clark, 2023).

We refer to this phenomenon as the **Duty to Due Diligence from Discoverable Documentation of Dangerous Deeds – or the 7D effec**t. Regulatory regimes that induce this effect are very helpful for improving accountability and reducing risks. Unfortunately, absent requirements for documentation and scrutiny thereof, developers in safety-critical fields have a perverse incentive to intentionally suppress documentation of dangers. For example, common legal advice warns companies against documenting dangers in written media:

> *These documents may not seem bad, but in the hands of an opposing attorney these cold hard facts can [be] used to swing a judge or a jury to their side. Often the drafters of these documents tend to believe that they are providing the company with some value to the business. For example, an engineer notices a potential liability in a design so he informs his supervisor through an email. However, the engineer's lack of legal knowledge... may later implicate the company... when a lawsuit arises.*
> – FindLaw Attorney Writers (2016), Safe Communication: Guidelines for Creating Corporate Documents That Minimize Litigation Risks

We personally enjoyed the use of "when" and not "if" in this excerpt.

Meanwhile, there is legal precedent for companies to sometimes lose court cases when they internally communicate risks through legally discoverable media such as in Grimshaw v. Ford (of Appeal, 1981). **So absent requirements, companies will tend to suppress the documentation of dangers to avoid accountability.** Meanwhile, mere voluntary transparency can be deceptive by selectively revealing information that reflects positively on the company (Ananny & Crawford, 2018). Thus, we argue that a regime that requires developers to produce scrutable documentation will be key to facilitating the production of more meaningful evidence.

4.4 CONSIDERING COUNTERARGUMENTS

"These 15 objectives would be too burdensome for developers." It's true that protocols and documentation can impose burdens. However, these burdens are generally far lighter than those imposed by substantive regulations, and compliance with many of these requirements may be trivial for developers already planning to take similar actions internally. For instance, even the most potentially burdensome measures – such as risk assessments and staged deployments – are practices that major developers like OpenAI, Anthropic, and Google DeepMind have already publicly committed to implementing.

"It's a slippery slope toward overreach." A second concern is that these 15 regulatory objectives might generate information that could pave the way for future substantive regulations or liability for developers. Regulatory and liability creep are real phenomena that can harm industry interests. However, it's important to emphasize that any progression from these objectives to future regulations or liability will ultimately depend on human decision-makers acting on evidence. Evidence is essential for society to engage in meaningful deliberation and exercise informed agency – this is the entire point of evidence-based policy. Therefore, if process-based AI regulations eventually lead to substantive regulations, it won't be because the process regulations laid an inevitable framework. It would be because the information produced by those regulations persuaded policymakers to take further action.

5 BUILDING A HEALTHIER ECOSYSTEM

Governing emerging technologies like AI is hard (Bengio et al., 2023). We don't know what is coming next. We echo the concerns of other researchers that there are critical uncertainties with the near and long-term future of AI. Anyone who says otherwise is probably trying to sell you something. So how do we go about governing AI under uncertainty? History teaches us some lessons about the importance of prescient action.

> *[Early] studies of global warming and the ozone hole involved predicting damage before it was detected. It was the prediction that motivated people to check for damage; research was intended in part to test their prediction, and in part to stimulate action before it was too late to stop...It was too soon to tell whether or not widespread and serious... damage was occurring, but the potential effects were troubling...A scientist would be in a bit of a bind: wanting to prevent damage, but not being able to prove that damage was coming...There are always more questions to be asked.*
> – Oreskes & Conway (2010), Merchants of Doubt

We often hear discussions about how policymakers need help from AI researchers to design technically sound policies. This is essential. But there is a two-way street. Policymakers can do a great deal to help researchers, governments, and society at large to better understand and react to AI risks.

Process regulations can lay the foundation for more informed debates and decision-making in the future. Right now, the principal objective of AI governance work is not necessarily to get all of the right substantive regulations in place. It is to shape the AI ecosystem to better facilitate the ongoing process of identifying, studying, and deliberating about risks. This requires being critical of the biases shaping the evidence we see and proactively working to seek more information. Kicking the can down the road for a lack of 'enough' evidence could impair policymakers' ability to take needed action.

This lesson is sometimes painfully obvious in retrospect. In the 1960s and 70s, a scientist named S.J. Green was head of research at the British American Tobacco (BAT) company. He helped to orchestrate BAT's campaign to deny urgency and delay action on public health risks from tobacco. However, he later split with the company, and after reflecting on the intellectual and moral irresponsibility of these efforts, he remarked:

> *Scientific proof, of course, is not, should not, and never has been the proper basis for legal and political action on social issues. A demand for scientific proof is always a formula for inaction and delay and usually the first reaction of the guilty. The proper basis for such decisions is, of course, quite simply that which is reasonable in the circumstance.*
> – S. J. Green, Smoking, Related Disease, and Causality

ACKNOWLEDGMENTS

We are thankful for our discussions with Akash Wasil, Ariba Khan, Aruna Sankaranarayanan, Dawn Song, Kwan Yee Ng, Landon Klein, Rishi Bommasani, Shayne Longpre, and Thomas Woodside.

REFERENCES

Preserving american dominance in artificial intelligence act of 2024. https://www.congress.gov/bill/118th-congress/senate-bill/5616/text, 2024. Accessed: 2025-02-04.

Mohamed Abdalla and Moustafa Abdalla. The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2020. URL https://api.semanticscholar.org/CorpusID:221995749.

Sara Abdulla and Husanjot Chahal. Voices of innovation: An analysis of influential ai researchers in the united states. July 2023. doi: 10.51593/20220022. URL https://doi.org/10.51593/20220022.

AI Index Steering Committee. Diversity in ai. In *The AI Index Report: Measuring Trends in Artificial Intelligence*, chapter 6. Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford, CA, 2021. URL https://aiindex.stanford.edu/ai-index-report-2021/.

Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20:973 – 989, 2018. URL https://api.semanticscholar.org/CorpusID:5001487.

Markus Anderljung, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. Towards publicly accountable frontier llms: Building an external scrutiny ecosystem under the aspire framework. *arXiv preprint arXiv:2311.14711*, 2023.

Anthropic. Responsible scaling program updates, October 2024. URL https://www.anthropic.com/rsp-updates. Accessed: 2024-11-21.

Sinan Arda. Taxonomy to regulation: A (geo) political taxonomy for ai risks and regulatory measures in the eu ai act. *arXiv preprint arXiv:2404.11476*, 2024.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*, 2024.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian K. Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atilim Gunes Baydin, Sheila A. McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Markus Brauner, and Sören Mindermann. Managing extreme ai risks amid rapid progress. *Science*, 384:842 – 845, 2023. URL https://api.semanticscholar.org/CorpusID:269929051.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, et al. International scientific report on the safety of advanced ai (interim report). *arXiv preprint arXiv:2412.05282*, 2024.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. International ai safety report. *arXiv preprint arXiv:2501.17805*, 2025.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. Recontextualizing fairness in nlp: The case of india. *arXiv preprint arXiv:2209.12226*, 2022.

Joseph Biden. Executive order 14110: Safe, secure, and trustworthy development and use of artificial intelligence, October 2023. URL https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. Accessed: 2024-11-21.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 173–184, 2022.

Rishi Bommasani, Sanjeev Arora, Yejin Choi, Li Fei-Fei, Daniel E. Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, Arvind Narayanan, Alondra Nelson, Emma Pierson, Joelle Pineau, Gaël Varoquaux, Suresh Venkatasubramanian, Ion Stoica, Percy Liang, and Dawn Song. A path for science- and evidence-based ai policy, 2024a. URL https://understanding-ai-safety.org/.

Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The foundation model transparency index v1.1: May 2024. *ArXiv*, abs/2407.12929, 2024b. URL https://api.semanticscholar.org/CorpusID:271270331.

Brazil. Bill No. 2338 of 2023: Regulating the Use of Artificial Intelligence, Including Algorithm Design and Technical Standards, 2023. URL https://digitalpolicyalert.org/event/11237-introduced-bill-no-2338-of-2023-regulating-the-use-of-artificial-intelligenc Accessed: 2024-11-21.

Paul Cairney. Evidence-based policymaking. In *European Commission Joint Research Centre*, pp. 1–3. 2021. URL https://paulcairney.wordpress.com/wp-content/uploads/2021/11/3_cairney_evidence-based-policymaking-16.11.21.pdf.

Canada. AI and Data Act: Part of Bill C-27, Digital Charter Implementation Act, 2022, 2022. URL https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading. Accessed: 2024-11-21.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Ben Bucknall, Andreas A. Haupt, Kevin Wei, J'er'emy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. URL https://api.semanticscholar.org/CorpusID:267301601.

Stephen Casper, Luke Bailey, Rosco Hunter, Carson Ezell, Emma Cabalé, Michael Gerovitch, Stewart Slocum, Kevin Wei, Nikola Jurkovic, Ariba Khan, et al. The ai agent index. *arXiv preprint arXiv:2502.01635*, 2025.

CASS. Model Artificial Intelligence Law Version 1.0 (Expert Suggestion Draft), 2023. URL https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-au Accessed: 2024-11-21.

Mauro Cazzaniga, Ms Florence Jaumotte, Longji Li, Mr Giovanni Melina, Augustus J Panton, Carlo Pizzinelli, Emma J Rockall, and Ms Marina Mendes Tavares. *Gen-AI: Artificial intelligence and the future of work*. International Monetary Fund, 2024.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 651–666, 2023.

China. Provisions on the Management of Algorithmic Recommendations in Internet Information Services, 2022a. URL https://www.chinalawtranslate.com/en/algorithms/. Accessed: 2024-11-21.

China. Provisions on the Administration of Deep Synthesis Internet Information Services, 2022b. URL https://www.chinalawtranslate.com/en/deep-synthesis/. Accessed: 2024-11-21.

China. Interim Measures for the Management of Generative Artificial Intelligence Services, 2023. URL https://www.chinalawtranslate.com/en/generative-ai-interim/. Accessed: 2024-11-21.

Sidney Dekker. *Foundations of Safety Science: A Century of Understanding Accidents and Disasters*. Routledge, Abingdon, UK and New York, USA, 2019. ISBN 978-1138493989.

Department of Commerce. Establishment of reporting requirements for the development of advanced artificial intelligence models and computing clusters. https://www.federalregister.gov/documents/2024/09/11/2024-20529/establishment-of-reporting-requirements-for-the-development-of-advanced-artificial 2024.

EU. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024. Accessed: 2024-11-21.

Chris Gebhardt. 1983-1986: The missions and history of space shuttle challenger. *NASA Space Flight*, 28, 2011.

Ronald N. Giere, John Bickle, and Robert F. Mauldin. *Understanding Scientific Reasoning*. Wadsworth Publishing, Belmont, CA, 5th edition, 2006. ISBN 978-0495004724.

Gillian K. Hadfield and Jack Clark. Regulatory markets: The future of ai governance. *ArXiv*, abs/2304.04914, 2023. URL https://api.semanticscholar.org/CorpusID:258060072.

Lennart Heim and Leonie Koessler. Training compute thresholds: Features and functions in ai regulation. *arXiv preprint arXiv:2405.10799*, 2024.

Dan Hendrycks and Thomas Woodside. A bird's eye view of the ml field [pragmatic ai safety #2]. *AI Alignment Forum*, 2022. URL https://www.alignmentforum.org/posts/AtfQFj8umeyBBkkxa/a-bird-s-eye-view-of-the-ml-field-pragmatic-ai-safety-2.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742*, 2024.

Sara Hooker. On the limitations of compute thresholds as a governance strategy. *ArXiv*, abs/2407.05694, 2024. URL https://api.semanticscholar.org/CorpusID:271051333.

Thomas S Kuhn. The structure of scientific revolutions. *University of*, 965, 1962.

Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, et al. A safe harbor for ai evaluation and red teaming. *arXiv preprint arXiv:2403.04893*, 2024.

Elliot McKernon, Gwyn Glasser, Deric Cheng, and Gillian Hadfield. Ai model registries: A foundational tool for ai governance. 2024. URL https://api.semanticscholar.org/CorpusID:273345835.

Shivani Metta, Isaac Chang, Jack Parker, Michael P. Roman, and Arturo F. Ehuan. Generative ai in cybersecurity, 2024. URL https://arxiv.org/abs/2405.01674.

Micah Musser. A cost analysis of generative language models and influence operations. *arXiv preprint arXiv:2308.03740*, 2023.

Arvind Narayanan and Sayash Kapoor. *AI Snake Oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press, 2024.

National Assembly of the Republic of Korea. Bill details: Prc_r2v4h1w1t2k5m1o6e4q9t0v7q9s0u0, 2025. URL https://likms.assembly.go.kr/bill/billDetail.do?billId=PRC_R2V4H1W1T2K5M1O6E4Q9T0V7Q9S0U0. Accessed: 2025-02-12.

S. Nevo, D. Lahav, A. Karpur, Y. O. G. E. V. Bar-On, H. A. Bradley, and J. Alstott. Securing ai model weights. Technical report, RAND Corporation, 2024. URL https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2800/RRA2849-1/RAND_RRA2849-1.pdf.

California Court of Appeal. Grimshaw v. ford motor co. 119, 1981. URL https://law.justia.com/cases/california/court-of-appeal/3d/119/757.html. Court decision concerning product liability and punitive damages related to the Ford Pinto.

OpenAI. Introducing the model spec, 2024. URL https://openai.com/index/introducing-the-model-spec/.

Naomi Oreskes and Erik M. Conway. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Publishing, 2010.

David Patterson. Technical perspective: For better or worse, benchmarks shape a field. *Commun. ACM*, 55(7), 2012.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*, 2022.

Rida Qadri, Renee Shelby, Cynthia L Bennett, and Emily Denton. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 506–517, 2023.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel E. Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022. URL https://api.semanticscholar.org/CorpusID:249605439.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H Kim, et al. Safetywashing: Do ai safety benchmarks actually measure safety progress? *arXiv preprint arXiv:2407.21792*, 2024.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 315–328, 2021.

Jonas B Sandbrink. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*, 2023.

Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K Hadfield, Richard Ngo, Konstantin Pilz, et al. Computing power and the governance of artificial intelligence. *arXiv preprint arXiv:2402.08797*, 2024.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*, 2024.

Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv preprint arXiv:2408.12622*, 2024.

Matthew Caleb Stephenson. Information acquisition and institutional design. *Harvard Law Review*, 124:1422–1484, 2011. URL https://api.semanticscholar.org/CorpusID:154434972.

Nassim Nicholas Taleb, Rupert Read, Raphael Douady, Joseph Norman, and Yaneer Bar-Yam. The precautionary principle (with application to the genetic modification of organisms). *arXiv preprint arXiv:1410.5787*, 2014.

USA. H.R. 9497, AI Advancement and Reliability Act, 2024a. URL https://science.house.gov/2024/9/h-r-xxxx-ai-advancement-and-reliability-act. Accessed: 2024-11-21.

USA. S.4178 - Future of Artificial Intelligence Innovation Act of 2024, 2024b. URL https://www.congress.gov/bill/118th-congress/senate-bill/4178. Accessed: 2024-11-21.

Yixin Wan and Kai-Wei Chang. White men lead, black women help: Uncovering gender, racial, and intersectional bias in language agency. *arXiv preprint arXiv:2404.10508*, 2024.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.