Position: Technical Research and Talent is Needed for Effective AI Governance

Anka Reuel [* 1]   Lisa Soder [* 2]   Ben Bucknall [3 4]   Trond Arne Undheim [1]

Abstract

In light of recent advancements in AI capabilities and the increasingly widespread integration of AI systems into society, governments worldwide are actively seeking to mitigate the potential harms and risks associated with these technologies through regulation and other governance tools. However, there exist significant gaps between governance aspirations and the current state of the technical tooling necessary for their realisation. In this position paper, we survey policy documents published by public-sector institutions in the EU, US, and China to highlight specific areas of disconnect between the technical requirements necessary for enacting proposed policy actions, and the current technical state of the art. Our analysis motivates a call for tighter integration of the AI/ML research community within AI governance in order to i) catalyse technical research aimed at bridging the gap between current and supposed technical underpinnings of regulatory action, as well as ii) increase the level of technical expertise within governing institutions so as to inform and guide effective governance of AI.

1. Introduction

The growing integration of artificial intelligence (AI) into various aspects of society over the past decade, including education, hiring, and finance, has became particularly noticeable with ChatGPT's release in late 2022. This event significantly heightened AI's visibility among both the general public and policymakers. The adoption of the technology, especially generative AI, has since surged across various sectors and is projected to further accelerate workforce automation and occupational switches (Ellingrud et al., 2023).

As the adoption of AI grows, so too does the awareness of its potential harms and risks. Over the past year, there has been a notable rise in reported AI-related incidents (Maslej et al., 2024)—i.e., instances where the failure or misuse of an AI system led to real-world harm, such as the production of harmful deepfakes or privacy breaches (McGregor, 2021). Given these developments, the need for effective, robust, and adaptable AI governance at national and international levels is more pressing than ever (Trager et al., 2023).

Governments around the world are responding to these challenges. For instance, Stanford's 2024 AI Index reveals a surge in AI-related laws, with relevant legislation increasing from one instance in 2016 to 28 by 2023 across 128 countries (Maslej et al., 2024). Among these developments, for example, China has recently enacted a new regulation on generative AI (China Law Translate, 2023b), while the EU has introduced the AI Act, with some rules scheduled to come into effect as early as Summer 2024 (European Commission, 2023). Likewise, the US has issued an Executive Order (E.O.) on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" and garnered voluntary commitments from major AI labs for governance measures (The White House, 2023c).

However, current gaps—both in terms of the technical tools and talent available to enact policy objectives—might negatively impact the capacity of governance actors to mitigate risks through regulation, potentially even lead to counterproductive outcomes (Guha et al., 2023). Poorly designed and executed policies, such as mandatory registration of AI systems, information disclosure requirements, or premarket approvals, could not only fail to ensure safe and reliable products but also burden AI research. Moreover, such policies risk creating entry barriers that disproportionately favor well-established, larger AI companies, potentially limiting competition and stifling innovation within the AI field (Dal Bó, 2006; Wu, 2023). To counter such issues, **we advocate for the following pair of positions:**

- **Targeted AI/ML research is urgently needed to ensure the effective enactment of current regulatory proposals.**

- **Closer collaboration between technical researchers and policy makers is necessary to ensure informed and effective governance.**

*Equal contribution [1]Department of Computer Science, Stanford University, Stanford, US [2]Interface, Brussels, BE [3]Centre for the Governance of AI, Oxford, UK [4]Oxford Martin AI Governance Initiative, Oxford, UK. Correspondence to: Anka Reuel <anka@cs.stanford.edu>.

The Need for Technical AI Governance

Overview of research areas & policy tasks that require further technical research & expertise

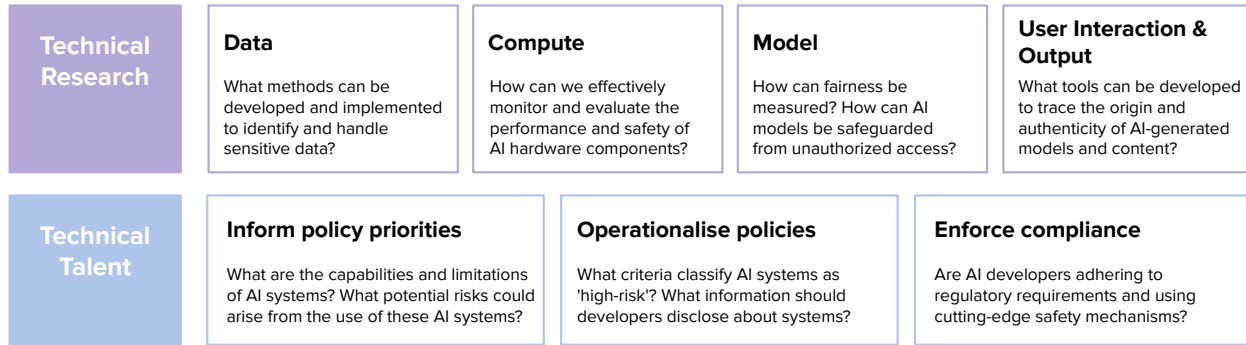| Technical Research	**Data** What methods can be developed and implemented to identify and handle sensitive data?	**Compute** How can we effectively monitor and evaluate the performance and safety of AI hardware components?	**Model** How can fairness be measured? How can AI models be safeguarded from unauthorized access?	**User Interaction & Output** What tools can be developed to trace the origin and authenticity of AI-generated models and content?
Technical Talent	**Inform policy priorities** What are the capabilities and limitations of AI systems? What potential risks could arise from the use of these AI systems?	**Operationalise policies** What criteria classify AI systems as 'high-risk'? What information should developers disclose about systems?	**Enforce compliance** Are AI developers adhering to regulatory requirements and using cutting-edge safety mechanisms?	

Figure 1. Overview of the role of technical research and expertise in AI governance. We propose that (i) targeted technical research is needed to address open tooling and research questions around data, compute, models, and deployment, and (ii) more technical expertise within governing institutions will be required to inform policy priorities, operationalise them into concrete policies, and enforce them effectively.

While advocating for these claims, we acknowledge that the governance of AI inevitably involves complex, value-laden decisions (Gordon et al., 2022; Kaminski, 2023), and reject techno-solutionist approaches to mitigating harms from AI. Addressing the manifold issues presented by such a technology necessitates a holistic and inter-disciplinary approach combining judicial, political, technological and broader societal insights, and we are consequently arguing for a closer collaboration between researchers and practitioners in these diverse and complementary fields.

This paper proceeds as follows. Section 2.1 presents a systematic review of policy documents published in the US, the EU, and China respectively, identifying the technical assumptions and tools upon which they rely. In Section 2.2, these requirements are contrasted against the current state of the art (SOTA) in AI/ML research, aiming to drawing attention to disparities between policy aspirations and technical realities. Section 2.3 suggests high-level research areas where further work could aid in addressing these gaps. Section 3 discusses the challenge posed by shortages of technical talent in governance institutions (Section 3.1) before highlighting opportunities for the AI/ML research community to assist in ongoing governance efforts (Section 3.2). Section 4 concludes.

2. Gap Between Policy Aspirations and Technical Research

Recent policy developments depend to a good extent on technical solutions that don't necessarily exist (yet). Often, policymakers set forth regulatory frameworks with the intention of ensuring, for instance, the safety, non-discrimination, and transparency in AI systems (Parliament & Council, 2024). However, these regulations sometimes depend on technological solutions that are not yet feasible given the current SOTA in AI/ML research. This mismatch implies that some regulatory requirements might be unattainable, highlighting a gap between regulatory expectations and technological capabilities. The gap necessitates further technical research to develop the tools and methodologies needed to meet and enforce these regulatory requirements.

In this chapter, we are providing concrete examples of this phenomenon from governance texts from the EU, US, and China, illustrating how certain policy prescriptions are outpacing the existing AI/ML research landscape, thereby emphasising the need for continued advancement and targeted, governance-supporting AI/ML research to bridge this divide. Finally, we provide an initial overview of a body of AI/ML research that the AI/ML community should prioritise to close the gap outlined above.

2.1. Relevant Regulatory Texts

For the purpose of this analysis, we will concentrate on AI legislation in the EU, US, and China. While AI governance frameworks proliferated rapidly across the globe in 2023, with 28 AI-related bills across 15 countries passing into law (Maslej et al., 2023), these three regions hold particular significance in the AI governance landscape. This is due to their influential roles shaping in AI research and development, as well as the potential for policy diffusion (Shipan & Volden, 2008; Bradford, 2020; Siegmann & Anderljung, 2022). This subsection provides an overview of the documents we're examining from each jurisdiction. The subsequent subchapter will highlight specific requirements

from these texts, pinpointing areas where they diverge from SOTA AI/ML research.

2.1.1. EUROPEAN UNION

Our focus of analysis in the EU is on the *European AI Act*, which is one of the first comprehensive pieces of legislation in this region directly focused on the regulation of AI systems (Parliament & Council, 2024). The act, which was first proposed by the EU Commission in 2021 (European Commission, 2022) and finally approved by the council in May 2024, is part of a broader legislative environment in the European Union, which includes the Digital Services Act, the Digital Markets Act, the General Data Protection Regulation, as well as national legislation of the member states (Hacker et al., 2024; Bogucki et al., 2022).

The **EU AI Act** (Parliament & Council, 2024) adopts a horizontal regulatory approach, meaning it establishes a common set of rules that apply uniformly across different sectors or industries. It follows a 'risk-based' approach, categorising AI systems into different levels of risk, each associated with a different set of regulatory obligations. High-risk systems, such as those used in critical infrastructures or for socio-economic decisions (e.g., hiring, education, financial services), face stringent requirements like conformity assessments, transparency obligations, and detailed record-keeping, while minimal-risk systems are largely free from regulatory obligations. GPAI systems have additional technical documentation requirements, and those posing systemic risks (e.g., models trained using compute greater than 10^{25} FLOPs) must conduct model evaluations, adversarial testing, and report serious incidents.

2.1.2. UNITED STATES

In the context of US AI governance, we focus on two key documents: the *Blueprint for an AI Bill of Rights* (The White House, 2022) and *Executive Order 14110* (The White House, 2023a). The **Blueprint for an AI Bill of Rights** (AIBoR), released on October 4, 2022, establishes a rights-based framework that outlines principles to guide the design, development, and deployment of automated systems (The White House, 2022). While the AIBoR is neither binding nor equips regulatory agencies with more enforcement power, we included it in our analysis as it provides a first indication of the approach in the Biden-Harris administration and already lays out resources and best practices to enact these principles.

Second, **E.O. 14110**, titled "Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (The White House, 2023a) and signed by President Biden on October 30, 2023, represents an initial approach by the US government to AI regulation. It mandates several federal agencies to appoint Chief AI Officers and

outlines specific responsibilities for agencies like the Department of Homeland Security, the Department of Veterans Affairs, and the National Institute of Standards and Technology (The White House, 2023a). The order has been described as the most comprehensive piece of governance by the US government pertaining to AI, following earlier initiatives such as the BAIBoR (Ryan-Mosley & Heikkilä, 2023).

2.1.3. CHINA

For China, we look at two different pieces of regulations: the *Provisions on the Management of Algorithmic Recommendations in Internet Information Services* (China Law Translate, 2021) and the *Provisions on the Administration of Deep Synthesis Internet Information Services* (China Law Translate, 2022)[1].

The **Provisions on the Management of Algorithmic Recommendations in Internet Information Services** (China Law Translate, 2021), introduced by China in 2021, is a regulatory framework designed to govern the use of algorithmic recommendation systems by internet service providers within the country. This legislation aims to ensure that algorithms promote positive content and uphold core socialist values, by requiring transparency in the operations of these algorithms and holding companies accountable for the content they recommend.

The **Provisions on the Administration of Deep Synthesis Internet Information Services** (China Law Translate, 2022), issued by China in 2022, represents a significant regulatory step towards overseeing what the Chinese call deep synthesis technologies, including deepfakes and other AI-generated content. This regulation aims to address the challenges and risks associated with the rapid development and application of deep synthesis in internet information services, particularly the potential for spreading misinformation, violating personal rights, and undermining social stability (China Law Translate, 2022).

2.2. Technical Requirements in AI Regulations

To structure the analysis of the documents, we focus on technical requirements across four areas where research and regulatory requirements intersect, based on a categorisation by (Bommasani et al., 2023): Data, Compute, Model, Output. Across these four categories, many of the techniques and tools required to effectively implement the proposed policies are either underdeveloped or non-existent. This disparity between regulatory requirements and the current state-of-the-art in AI/ML research poses significant chal-

---

[1]Given the limited scope of the paper, we omitted the *Interim Measures for the Management of Generative Artificial Intelligence Services* (China Law Translate, 2023a)

lenges to the effective governance of AI systems. The following section illuminates specific examples of these gaps and their implications for AI governance.

2.2.1. DATA

All three regions, the EU, the US, and China, have data governance requirements in the regulatory documents outlined in Section 2.1. For example, the EU AI Act states:

> *"The data sets should also have the appropriate statistical properties, including as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used, with specific attention to the mitigation of possible biases in the data sets."* (Parliament & Council, 2024, Recital 67)

Technical bias identification and mitigation within AI systems comes with several challenges: For example, both necessitate a clear definition of fairness – a concept that remains inherently ambiguous and subjective in the context of AI (Mehrabi et al., 2021; Pagano et al., 2023). Despite the critical importance of ensuring equitable outcomes in AI applications, there is currently no universally accepted framework or structured approach for determining the appropriate technical implementation of fairness.

In the US, the E.O.and the BAIBoR both outline requirements for data governance and data privacy:

> *"Artificial Intelligence's capabilities [...] can increase the risk that personal data could be exploited and exposed. To combat this risk, the Federal Government will ensure that the collection, use, and retention of data is lawful, is secure, and mitigates privacy and confidentiality risks. Agencies shall use available policy and technical tools, including privacy-enhancing technologies (PETs) where appropriate, to protect privacy [...]"* (The White House, 2023a, Section 2(f))

A similar requirement is put forth in the Chinese AI regulations:

> *"Deep synthesis service providers shall implement primary responsibility for information security, establishing and completing management systems such as for [...] data security, personal information protection [...] and shall have safe and controllable technical safeguard measures."* (China Law Translate, 2022, Article 7)

The possibility for protection and guarantee of privacy vary depending on the type of AI system in use. For instance, LLMs have been found vulnerable to revealing personally identifiable information (PII) under adversarial attacks, such as through adversarial prompting strategies (Nasr et al., 2023; Ippolito et al., 2023). Currently, there is no robust method to safeguard privacy in these cases. Similar privacy breaches have also been observed in non-generative models (Rigaki & Garcia, 2023; Cristofaro, 2020).

> *"You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used."* (The White House, 2022)

For users to have agency, they need to be informed about if and how their data was utilised in a model. The provenance of training data in AI systems, particularly in LLMs, is often undisclosed (Longpre et al., 2023; Ojewale et al., 2024). At the same time, there are only limited possibilities to find out, without white-box access to the model, what was in the training data, making it difficult for users to know if and for what their data was used if it has not been proactively been disclosed by the model provider (Longpre et al., 2023; 2024b; Casper et al., 2024). Moreover, attributing training data is difficult, especially without access to the model, and recent advancements in the field aren't applicable to all model classes (Park et al., 2023). Therefore, even if users are aware that their data was included in the training set, they rarely understand its influence on the model's behaviour, which further limits their agency.

2.2.2. COMPUTE

Both the US and the EU have registration requirements for model developers that train large models; the EU defined a corresponding FLOPs threshold beyond which a model is considered to have systemic risk (and hence is subject to more scrutiny):

> *"A general-purpose AI model shall be presumed to have high impact capabilities pursuant to paragraph 1, point (a) [definition of systemic risk], when the cumulative amount of computation used for its training measured in floating point operations is greater than $10(^25)$."* [Art 51](Parliament & Council, 2024)

> *"Where a general-purpose AI model meets the condition referred to in Article 51(1), point (a), [referring to GPAI systems with systemic risks] the relevant provider shall notify the Commission without delay [...]"* (Parliament & Council, 2024, Article 52, 1)

A related requirement can be found in the US E.O., that requires

> *"Companies developing or demonstrating an intent to develop potential dual-use foundation models to provide the Federal Government, on an ongoing basis, with information, reports, or records regarding [...] (A) any ongoing or planned activities related to training, developing, or producing dual-use foundation models[...]"* (The White House, 2023a, Sec. 4.2(i))

Although this requirement is not technically impossible to fulfil, enforcing it will prove challenging. Currently, there's no method to verify if a large model is being trained. Some proposed solutions include some form of on-chip governance mechanisms and chip control, e.g., using chips that either limit or report the number of computations they perform (Aarne et al., 2024).

2.2.3. MODELS

All regions we analysed put forward requirements for evaluating, reporting, and mitigating risks from AI systems. For example, in the EU AI Act requires

> *"providers of general-purpose AI models with systemic risk shall: (a) perform model evaluation in accordance with standardised 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 risks."* (Parliament & Council, 2024, Article 55)

Whereas the E.O.in the US has similar evaluation requirements for 'dual-use foundation models':

> *"(i) Companies developing or demonstrating an intent to develop potential dual-use foundation models to provide the Federal Government, on an ongoing basis, with information, reports, or records regarding the following: [...] (C) the results of any developed dual-use foundation model's performance in relevant AI red-team testing based on guidance developed by NIST pursuant to subsection 4.1(a)(ii) of this section, and a description of any associated measures the company has taken to meet safety objectives, such as mitigations to improve performance on these red-team tests and strengthen overall model security."* (The White House, 2023a, Section 4.2 (C))

Although numerous jurisdictions mandate capability evaluations across a variety of risk areas, from a technical stand-

point, there is a lack of clarity on how to perform these assessments both comprehensively and reliably (Chang et al., 2023; Zhou et al., 2023). Weidinger et al. (2023), in their review of current safety evaluation practices, noted that for a wide range of risk scenarios, evaluations simply do not exist yet. Even for risk areas where evaluations are available, there is still a substantial degree of uncertainty regarding their ability to accurately capture and assess specific concepts (Raji et al., 2021; Liao & Xiao, 2023), and as such how much trust to put into evaluations (Liao et al., 2021; Zhou et al., 2023; Mitchell, 2023).

Furthermore, the US included access and security requirements for model weights in their E.O.:

> *"(i) Companies developing or demonstrating an intent to develop potential dual-use foundation models to provide the Federal Government, on an ongoing basis, with information, reports, or records regarding the following: [...] (B) the ownership and possession of the model weights of any dual-use foundation models, and the physical and cybersecurity measures taken to protect those model weights"* (The White House, 2023a, Section 4.2 (B))

Securing models from leakage has proven to be difficult (Franzen, 2024; Vincent, 2023). Some technical ideas to de-incentivise leackage of models has been to watermark models so that it can be traced back to who released a model (Mehta et al., 2022; Boenisch, 2021). Alternative ideas included self-destructing models (Henderson et al., 2023), but both are underexplored and non-production-ready ideas that would need to be investigated further.

2.2.4. USER INTERACTION AND OUTPUT

Finally, the US mentions consumer protection from harmful, ineffective models:

> *"You should be protected from unsafe or ineffective systems."* (The White House, 2022)

Protection from ineffective or unsafe systems is difficult to ensure. One issue is that no robust evaluation methods exist to comprehensively understand whether a model is harmless or not (see Section 2.2.3). Furthermore, AI systems that are released may be ineffective. For example, in the context of generative AI, no robust technical approach to outputting hallucination-free, reliable information exists yet, which renders generative AI systems ineffective for many use cases (Rawte et al., 2023).

Similarly, regulation from China stipulates that the output of a system has to be controlled:

[...] persist [AI systems] in being oriented towards mainstream values, optimize mechanisms for algorithmic recommendation services, actively transmit positive energy, and promote the uplifting use of algorithms. The providers of algorithmic recommendation services [...] shall employ measures to prevent and stop the transmission of negative information (China Law Translate, 2021)

While approaches such as Constitutional AI (Bai et al., 2022) exist to somewhat align models with certain values, these approaches are far from perfect and do not ensure adherence to the predetermined values all the time (Ji et al., 2023). Additionally, it has been shown that these guardrails can be circumvented (Wolf et al., 2023; Zhang et al., 2023).

To reduce risks posed by synthetic content, the US E.O. lays out priorities to identify current practices, as well as support the potential development of

"further science-backed standards and techniques, for (i) authenticating content and tracking its provenance; (ii) labeling synthetic content, such as using watermarking; (iii) detecting synthetic content;" (The White House, 2023b, Section 4.5 (a))

While there have been significant advancements in watermarking AI-generated content (Zhao et al., 2023; Sadasivan et al., 2023; Ghosal et al., 2023; Fröhling & Zubiaga, 2021), the detection of non-watermarked, AI-generated content has proven to be difficult for all modalities (Li et al., 2023; Weber-Wulff et al., 2023; Heidari et al., 2023). Especially in light of advancements in generated video, image, and audio quality, detection of AI-generated content will remain challenging, necessitating more research in this area to enact these regulations.

2.3. Prioritising AI/ML Research for AI Governance

We define technical research for AI governance as *the development and deployment of technical tools, methods, and techniques for supporting effective AI governance, as well as analysis that motivates and informs their implementation.* Based on the comparative analysis and the gaps identified, as well as following from our definition, we suggest to prioritise AI/ML research addressing technical shortcomings in the four categories of *data*, *compute*, *models*, and *deployment* presented above. Figure 2 provides example high-level topics within each of these categories that could be addressed by future research.[2]

---

[2]A paper on open technical problems in AI governance by two of this papers' authors, Anka Reuel and Ben Bucknall, is forthcoming.

Examples of Open Research Topics in Technical AI Governance

Data	• Developing methods for identifying sensitive, copyrighted or harmful data in training, fine-tuning, or retrieval datasets • Proposing methods for detecting or preventing the extraction of training data from AI systems
Compute	• Designing methods for differentiating between AI chip workloads (e.g. training vs. inference) based on chip metadata • Development and application of trusted execution environments on AI chips
Models	• Improving the robustness and reliability of metrics and evaluations of AI systems • Proposing methods for providing secure researcher and auditor access to AI models
User Interaction & Output	• Developing of robust tooling for determining the provenance of AI-generated content • Designing methods for evaluating the downstream societal impacts of AI systems

Figure 2. Technical Research: Examples of areas that require further technical research to allow for supporting effective AI governance. These are not meant to be exhaustive lists. We also want to note that for most of the examples highlighted, there already exists research to a varying degree. However, none of the areas are sufficiently well studied, as outlined in Section 2.2, to close the gaps highlighted in Section 2.2, and hence warrant more attention from the AI/ML research community.

3. AI Governance Requires More Technical Expertise

A core question at the heart of any regulatory regime – and particularly in those as complex and rapidly evolving as the AI domain – is whether the regulator is equipped with the expertise and capacity necessary (Weissinger, 2022; Baldwin et al., 2011) to i) inform and guide governance priorities, ii) operationalise these priorities into concrete policies and verifiable requirements and iii) enforce them. Building on the previous sections' focus on the technical tools needed to enact policies, we now turn our attention to the expertise required to develop an effective AI accountability infrastructure and governance capacity. In this section, we will first highlight some of the technical talent shortages AI governance faces. We will then elaborate on the need for AI/ML researchers to support the key tasks along the policy cycle. Finally, we will provide a brief overview and examples of avenues for ML researchers to take a more active role in AI governance efforts.

3.1. Technical Talent Gap in AI Governance

Currently, there is little empirical evidence on the preparedness of governments to effectively govern AI technologies. However, graduation data in the US paints a concerning

picture, with less than 1% of AI PhD graduates choosing government careers post-study, while over 65% enter the private sector and more than 25% pursue academic careers (Zweben & Bizot, 2017). Drawing a similar conclusion, the need for more technical experts in AI governance has been raised as a key concern and priority (Guha et al., 2023; Schmidt et al., 2021; Engstrom et al., 2020). This challenge is not unique to the US; similar worries about attracting sufficient technical talent have been voiced in other jurisdictions, including the UK (Aitken et al., 2022) and EU member countries like Germany (Engler, 2022).

To address some of these capacity constraints, regulators often outsource some regulatory activities to external, private sector efforts, i.e., "regulatory intermediaries" (Abbott et al., 2017) such as standard-setting bodies and third-party auditors. While such an approach can offer more flexibility - for instance, private organisations such as AI auditors might be able to offer higher salaries than traditional government roles, these approaches, too, suffer from their own capacity shortages. For instance, at present, the ecosystem of third-party research organisations equipped to perform capability evaluations or risk assessments is still underdeveloped (Engler, 2023). Additionally, standard-setting bodies, while able to harness technical expertise from industry (Blind et al., 2023; Ramanna, 2015), frequently fall short in integrating a broader and more independent range of expertise to tackle the socio-technical nature of their work (Edwards, 2022).

3.2. The Need for Technical Expertise in AI Governance

In this final chapter, we highlight three different tasks across the policy lifecycle where technical researchers can contribute to supporting AI governance efforts.

Examples of Tasks Along the Policy Cycle That Require Technical Expertise

Inform Priorities	• Monitoring and communicating key trends in AI development and deployment • Evaluating AI systems to better understand current capabilities and real-world impacts
Operationalise Policies	• Establishing criteria for the risk classification of AI systems • Developing guidelines for model reporting, information sharing and safety testing
Enforce Compliance	• Conducting AI system audits and conformity assessments • Advising courts on interpreting technical evidence in legal proceedings

Figure 3. Technical Talent: Examples of tasks along key steps in the policy cycle that will require technical expertise.

3.2.1. INFORM AND GUIDE POLICY PRIORITIES

A key challenge for AI governance is to set the right priorities and design targeted policies that neither fall short in their mandate to protect the public interest nor become excessively intrusive and hamper innovation. To achieve this goal, policymakers will need to have the necessary means to source reliable information about the progress of AI systems, their limitations, and potential future trajectories (Whittlestone & Clark, 2021; Bengio et al., 2024a; Clark, 2024; Coglianese et al., 2004). Considering the influential role of information in setting policy agendas (Jones & Baumgartner, 2005), ensuring that policies align with the public interest will necessitate integrating diverse information sources beyond industry, the AI/ML research community being one of them.

Drawing a parallel to climate regulation and the Intergovernmental Panel on Climate Change (IPCC), scientists play an important role in informing government decisions by providing climate models and a body of evidence to base decisions on further developments, and AI governance will need similar mechanism to guide policy making.

One such example within the AI governance space, where we've already observed a government body aiming to build the infrastructure to deliver such information, is the UK AI Safety Institute (AISI). The UK's AISI was tasked with building *"a body of evidence on the risks from advanced AI [...] to lay the foundations for technically grounded international governance."* (Department for Science, Innovation and Technology, 2023). In working towards this aim, the AISI intends to carry out three main activities: developing and conducting evaluations on advanced AI systems, driving foundational AI safety research (including *"building products for AI governance"* (Department for Science, Innovation and Technology, 2023)), and facilitating information exchange.

Several other jurisdictions, including the US (NIST, 2023), Japan (Shimbun, 2023) and Canada (Department of Finance Canada, 2024), have followed suit, announcing initiatives akin to the AISI. Furthermore, the European Commission has announced an independent expert panel in which AI researchers could advise on the governance of general-purpose AI systems in a scientific capacity (European Commission, 2023).

3.2.2. OPERATIONALISE POLICIES

Once policy priorities and interventions are formulated, a key challenge lies in translating these into more concrete obligations and verifiable claims. This step can encompass tasks such as defining the scope and character of activities to be integrated into risk management processes (Ozlati & Yampolskiy, 2017; Schuett, 2023; Pouget & Zuhdi, 2024),

outlining reporting requirements (Kolt et al., 2024; Bommasani et al., 2024) or establishing the criteria of what constitutes an AI system of "systemic" or "high" risk (Edwards, 2022; Moës & Ryan, 2023).

Defining different risk levels, for example, will be a key factor, both in the US and EU, in determining the level of scrutiny an AI system should undergo (The White House, 2023a; European Commission, 2022). However, establishing these demands a deep understanding of the current capabilities and limitations of AI systems. A lack thereof can have significant consequences, as regulatory failures in other industries have previously demonstrated. For instance, in the case of the Boeing 737 groundings, while multiple factors contributed to the regulatory failure, such as limited resources and industry pressure, a key issue was that Federal Aviation Administration engineers did not fully grasp the risk profile of the Maneuvering Characteristics Augmentation System (MCAS), the software implicated in the crashes (Kitroeff et al., 2019; DeFazio & Larsen, 2020). This lack of understanding of the underlying technology led to the software being erroneously categorised as low-risk, thereby bypassing more thorough scrutiny.

Beyond the mere availability of expertise, it is also crucial to consider its sources. In order to operationalise policy goals, most AI regulatory regimes reference to standards (Pouget, 2023). These are usually developed by standard-setting organisations, which in turn predominantly rely on contributions from the very industry it seeks to regulate. Such reliance is generally less of an issue in scenarios where the goals of the industry and regulators are closely aligned, like in ensuring the interoperability of electronic devices. However, the situation becomes more precarious when standard-setting organisations directly shape the regulatory obligations of developers. Put differently, regulatory requirements can have a direct impact on a developer's bottom line, creating the potential for regulatory capture (Dal Bó, 2006). Given this dynamic there's a need for greater involvement of the AI/ML research community (as well as other groups, such as representatives from civil society (Edwards, 2022)) to ensure standards are well-informed and drive forward more robust and equitable regulatory practices (Cihon, 2019).

3.2.3. ENFORCE COMPLIANCE

After establishing policy priorities and converting them into tangible requirements, another crucial step in ensuring that rules will effectively bring about change lies in the regulator's capacity to enforce these regulations. While, in the first section, we mostly elaborated on the technical tools underlying enforcement, there is also a key question towards the people, infrastructure and expertise necessary to hold AI developers accountable (Scherer, 2015). Without a good

understanding of the underlying practices in AI development, as well as the technology more generally, it will be difficult to detect violations and, for instance, scrutinise claims made by AI developers. Furthermore, narrowing information asymmetries not only increases the likelihood of identifying infringements but also strengthens overall compliance by presenting a more credible threat of detection and action (Ayres & Braithwaite, 1992; Baldwin & Black, 2008).

A concrete example where the expertise of the AI/ML research community could be crucial is in the context of the EU Liability Directive. While at the time of writing, this legislative act is still in the proposal stage, AI developers could be held liable for harms that are 1) foreseeable and 2) a likely result from a failure to meet "duty of care" (European Parliament & European Council, 2022, Article 4). In essence, if a developer neglects established best practices and standards in AI system development, and this oversight leads to foreseeable harm, they could be liable (Li et al., 2022). To investigate a claim, the current proposal would grant the European Commission powers to request technological evidence and documentation. However, determining misconduct might largely hinge on the regulator's ability to scrutinise the information provided by the industry and to ascertain whether the conditions of predictability (i.e., foreseeability of harm) and negligence of the duty of care (i.e., adherence to best practices) were met. Given that AI development is fundamentally grounded in practices from the AI/ML research community, the community is uniquely positioned to bridge some of the information gaps to hold AI developers accountable.

Alongside expertise housed within regulatory bodies and policymaking institutions, the effective enforcement of current AI regulations critically relies on the availability of independent, third-party auditing and testing services for external scrutiny (Hadfield & Clark, 2023; Raji et al., 2022; Anderljung et al., 2023). The responsibilities of such organisations are likely to include tasks such as red-teaming of AI systems on safety properties, or assessing the conformity with certain standards. While, from a policy perspective, there are still numerous uncertainties, such as the exact scope of audits and the processes for appointing and certifying auditors (Ojewale et al., 2024; Guha et al., 2023), a primary challenge, to begin with, is the limited availability of such organisations (Engler, 2023).

3.2.4. OPPORTUNITIES TO WORK ON AI GOVERNANCE

While there are numerous ways the AI/ML research community can contribute to AI governance efforts—more than we could exhaustively capture here—below we provide an overview of some key avenues that require varying levels of involvement:

Part-time Opportunities

- **Technical AI Governance Research**: This could involve developing new insights into open technical questions as outlined above, collaborating with interdisciplinary scholars in fields such as sociology, health, or law, or synthesising existing information for broader audiences, similar to the AI Index (Maslej et al., 2024) or the International Scientific Report on AI Safety (Bengio et al., 2024b).

- **Policy Advisory Roles**: Another way to enhance government expertise without fully leaving academia is by taking on advisory roles, for instance by serving on the EU's AI scientific panel (Council, 2024), the UN's Multistakeholder Advisory Body on AI (UN Technology Envoy, 2023), or taking sabbaticals to contribute to policy development (Dworkin, 2024). Other lower-commitment options that are open to all are participation in working groups or responding to requests for comments on governmental AI governance initiatives.

- **Standards Setting**: Researchers can shape AI governance and industry practices by contributing to national standards-setting bodies (e.g., NIST, CEN, CEN-ELEC), which often detail regulatory provisions, or private standards organizations (e.g., ISO/IEC, IEEE) that are frequently referenced by national bodies and become common practice in the industry.

Full-time Opportunities

- **Policy practice**: Recently, governments have placed a strong emphasis on recruiting technical talent, as evidenced by initiatives such as the National AI Talent Surge in the US (White House OSTP, 2024) or the focus on technical expertise in the first hiring round of the EU AI office (2024). Furthermore, fellowship programs, including TechCongress (TechCongress, 2024) and Horizon (Horizon, 2024) in the US, can provide supported pathways into policy roles.

- **Think Tanks**: Conducting research at a think tank could involve translating technical insights into policy recommendations, mapping the landscape of AI developments, similar to Epoch's efforts in mapping AI trends (Sevilla et al., 2022), or reviewing the effectiveness of current policies, such as Groves et. al (2024) work on reviewing algorithmic bias audit regime.

- **Regulatory Entrepreneurship**: Another opportunity to shape AI governance is by building organisations that provide the necessary AI accountability infrastructure, e.g., by developing privacy-enhancing technologies similar to OpenMined's work (2020) or providing evaluation and auditing services.

We believe that technical AI governance offers many exciting opportunities. However, we also acknowledge that pursuing these paths may come with considerable costs for AI/ML researchers, such as significantly lower public sector salaries compared to industry role (U.S. Bureau of Labor Statistics, 2023; Fleming, 2023) or slower progress towards tenure when undertaking additional responsibilities.

Addressing these issues will require additional initiatives from other stakeholders in the AI governance ecosystem, such as universities, governments, and industry, to incentivize more technical AI governance work and interdisciplinary collaboration. While a thorough exploration of potential solutions exceeds this paper's scope, examples of such initiatives could include funding national interdisciplinary research programs, addressing pay disparities, ensuring legal protections for AI auditors (Longpre et al., 2024a), offering and promoting policy fellowships, or providing more opportunities for cross-disciplinary networking.

4. Conclusion

In this position paper we have argued for closer integration between AI researchers and policy makers in governments and other public-sector bodies. We surveyed policy documents published in the EU, US, and China, finding significant divergence between the current state of technical solutions and those assumed by, and necessary for enacting, proposed policy actions. In particular, we focused on technical requirements at the *data*, *compute*, *model*, and *deployment* level. Furthermore, we argued that government bodies will require more access to relevant technical expertise for *informing*, *operationalising*, and *enforcing* governance and policy actions.

Based on these arguments we make the following two calls to action to the AI/ML research community:

1. To prioritise research topics targeted at narrowing the gap between assumed and actual technical tools available for supporting governance efforts; and

2. To work towards a closer integration with policy-makers, so as to ensure informed and effective governance of AI.

We look forward to following and supporting work towards these two aims, and plan on publishing a research agenda for the former as a next step in the near future.

Impact Statement

With this position paper, we seek to motivate the AI research community to actively engage with AI governance efforts, ultimately aiming to increase governance capacity to anticipate and respond to challenges posed by AI systems. We argue that AI researchers can contribute to this goal in two key ways: first, by creating and advancing the research necessary for effective oversight, and second, by providing their expertise to inform, operationalise and enforce policies. However, the technical research directions we propose only serve as illustrative examples rather than a comprehensive roadmap or a prioritisation of the most critical issues. Moreover, technical solutions and expertise are only one component of the AI governance toolbox; effective AI governance will require insights from other areas such as the social sciences, public engagement, and other interdisciplinary perspectives—areas that fall beyond the scope of this paper.

References

Aarne, O., Fist, T., and Withers, C. Secure, Governable Chips, January 2024. URL `https://www.cnas.org/publications/reports/secure-governable-chips`. Accessed: 2024-1-28.

Abbott, K. W., Levi-Faur, D., and Snidal, D. Introducing regulatory intermediaries. *The Annals of the American Academy of Political and Social Science*, 670(1):6–13, 2017.

Aitken, M., Leslie, D., Ostmann, F., Pratt, J., Margetts, H., and Dorobantu, C. Common Regulatory Capacity for AI. *The Alan Turing Institute.*, 2022.

Anderljung, M., Smith, E., O'Brien, J., Soder, L., Bucknall, B., Bluemke, E., Schuett, J., Trager, R., Strahm, L., and Chowdhury, R. Towards publicly accountable frontier LLMs. In *Socially Responsible Language Modelling Research*, 2023.

Ayres, I. and Braithwaite, J. *Responsive regulation: Transcending the deregulation debate*. Oxford University Press, USA, 1992.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Baldwin, R. and Black, J. Really responsive regulation. *The modern law review*, 71(1):59–94, 2008.

Baldwin, R., Cave, M., and Lodge, M. What is 'Good' Regulation? In *Understanding Regulation: Theory, Strategy, and Practice*. Oxford University Press, 10 2011.

URL `https://doi.org/10.1093/acprof:osobl/9780199576081.003.0003`.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J., and Mindermann, S. Managing extreme AI risks amid rapid progress. *Science*, 384(6698):842–845, 2024a. doi: 10.1126/science.adn0117. URL `https://www.science.org/doi/abs/10.1126/science.adn0117`.

Bengio, Y., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Goldfarb, D., Heidari, H., Khalatbari, L., Longpre, S., Mavroudis, V., Mazeika, M., Ng, K. Y., Okolo, C. T., Raji, D., Skeadas, T., Tramèr, F., and Mindermann, S. International scientific report on the safety of advanced ai interim report. Technical report, UK Department for Science, Innovation and Technology, May 2024b. URL `https://assets.publishing.service.gov.uk/media/66474eab4f29e1d07fadca3d/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf`.

Blind, K., Kenney, M., Leiponen, A., and Simcoe, T. Standards and innovation: A review and introduction to the special issue. *Research Policy*, 52(8):104830, 2023.

Boenisch, F. A systematic review on model watermarking for neural networks. *Frontiers in Big Data*, 4:729663, 2021.

Bogucki, A., Engler, A., Perarnaud, C., and Renda, A. The AI Act and emerging EU digital acquis. *CEPS (Centre for European Policy Studies).*, 2022. URL `https://www.ceps.eu/ceps-publications/the-ai-act-and-emerging-eu-digital-acquis`.

Bommasani, R., Klyman, K., Zhang, D., and Liang, P. Do Foundation Model Providers Comply with the EU AI Act?, 2023. URL `https://crfm.stanford.edu/2023/06/15/eu-ai-act.html`.

Bommasani, R., Klyman, K., Longpre, S., Xiong, B., Kapoor, S., Maslej, N., Narayanan, A., and Liang, P. Foundation Model Transparency Reports. *arXiv preprint arXiv:2402.16268*, 2024.

Bradford, A. *The Brussels effect: How the European Union rules the world*. Oxford University Press, USA, 2020.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Hagen, M. V., Alberti, S.,

Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-Box Access is Insufficient for Rigorous AI Audits. *arXiv preprint arXiv:2401.14446*, 2024.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

China Law Translate. Provisions on the Management of Algorithmic Recommendations in Internet Information Services. *China Law Translate*, Dec 2021. URL https://www.chinalawtranslate.com/en/algorithms/. Accessed: 2024-01-23.

China Law Translate. Provisions on the Administration of Deep Synthesis Internet Information Services. *China Law Translate*, Nov 2022. URL https://www.chinalawtranslate.com/en/algorithms/. Accessed: 2024-01-23.

China Law Translate. Interim Measures for the Management of Generative Artificial Intelligence Services. *China Law Translate*, Aug 2023a. URL https://www.chinalawtranslate.com/en/generative-ai-interim/.

China Law Translate. Interim Measures for the Management of Generative Artificial Intelligence Services. *China Law Translate*, July 2023b. URL https://www.chinalawtranslate.com/en/generative-ai-interim/. Accessed: 2024-1-31.

Cihon, P. Standards for AI governance: international standards to enable global coordination in AI research & development. *Future of Humanity Institute, University of Oxford*, pp. 340–342, 2019.

Clark, J. Information Markets and AI Development. *The Oxford Handbook of AI Governance*, pp. 345, 2024.

Coglianese, C., Zeckhauser, R., and Parson, E. Seeking truth for power: Informational strategy and regulatory policymaking. *Minn. L. Rev.*, 89:277, 2004.

Council. Artificial Intelligence (AI) Act: Council gives final green light to the first worldwide rules on AI. Press Release, May 2024. URL https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/#:~:text=The%20new%20law%20aims%20to,on%20artificial%20intelligence%20in%20Europe.

Cristofaro, E. D. An Overview of Privacy in Machine Learning. *arXiv preprint arXiv:2005.08679*, 2020.

Dal Bó, E. Regulatory capture: A review. *Oxford review of economic policy*, 22(2):203–225, 2006.

DeFazio, P. A. and Larsen, R. Final Committee Report on the Design, Development, & Certification of the Boeing 737 MAX. Technical report, Committee on Transportation and Infrastructure, September 2020. Prepared by Majority Staff of the Committee on Transportation and Infrastructure.

Department for Science, Innovation and Technology. Introducing the AI Safety Institute, November 2023. URL https://www.gov.uk/government/publications/ai-safety-institute-overview. Accessed: 2024-2-1.

Department of Finance Canada. Remarks by the Deputy Prime Minister on securing Canada's AI advantage, April 2024. URL https://www.canada.ca/en/department-finance/news/2024/04/remarks-by-the-deputy-prime-minister-on-securing-canadas-ai-advantage.html.

Dworkin, J. How to boost your research: take a sabbatical in policy. *Nature*, 626(8000):694–694, 2024.

Edwards, L. Regulating AI in Europe: four problems and four solutions. *Ada Lovelace Institute*, 15:2022, 2022. URL https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf.

Ellingrud, K., Sanghvi, S., Dandona, G. S., Madgavkar, A., Chui, M., White, O., and Hasebe, P. Generative AI and the future of work in America, July 2023. URL https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america. Accessed: 2024-1-31.

Engler, A. Institutionalizing Data Analysis in German Federal Governance. *Brookings Institution*, 2022.

Engler, A. Key enforcement issues of the AI Act should lead EU trilogue debate. *Brookings Institution*, 2023.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., and Cuéllar, M.-F. Government by algorithm: Artificial intelligence in federal administrative agencies. *NYU School of Law, Public Law Research Paper*, (20-54), 2020.

European Commission. A European approach to artificial intelligence, 2022. URL https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence. Accessed: 2024-01-23.

European Commission. Commission welcomes political agreement on Artificial Intelligence Act, December 2023. URL https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473. Accessed: 2024-2-1.

European Commission. Job opportunities at the European AI Office. https://digital-strategy.ec.europa.eu/en/news/job-opportunities-european-ai-office, 3 2024. NEWS ARTICLE, Published 07 March 2024.

European Parliament and European Council. Directive on Adapting Non-contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive). Policy Document COM(2022) 496 final, European Union, September 28 2022. Proposal for a Directive of the European Parliament and of the Council.

Fleming, A. Government Tech Workers Earn Less. Is That a Bad Thing? Web Page, 2023. URL https://citizencodex.com/our-stories/government-tech-workers-earn-less-is-that-a-bad-thing.

Franzen, C. Mistral CEO confirms 'leak' of new open source AI model nearing GPT-4 performance, 01 2024. URL https://venturebeat.com/ai/mistral-ceo-confirms-leak-of-new-open-source-ai-model-nearing-gpt-4-performance/. Accessed: 2024-02-01.

Fröhling, L. and Zubiaga, A. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7:e443, 2021.

Ghosal, S. S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., and Bedi, A. A Survey on the Possibilities & Impossibilities of AI-generated Text Detection. *Transactions on Machine Learning Research*, 2023.

Gordon, G., Rieder, B., and Sileno, G. On mapping values in AI governance. *Computer Law & Security Review*, 46: 105712, 2022.

Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., and Strait, A. Auditing Work: Exploring the New York City algorithmic bias audit regime. *arXiv preprint arXiv:2402.08101*, 2024.

Guha, N., Lawrence, C., Gailmard, L. A., Rodolfa, K., Surani, F., Bommasani, R., Raji, I., Cuéllar, M.-F., Honigsberg, C., Liang, P., et al. AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review, Forthcoming*, 2023.

Hacker, P., Cordes, J., and Rochon, J. Regulating Gatekeeper Artificial Intelligence and Data: Transparency, Access and Fairness under the Digital Markets Act, the General Data Protection Regulation and Beyond. *European Journal of Risk Regulation*, 15(1):49–86, 2024.

Hadfield, G. K. and Clark, J. Regulatory Markets: The Future of AI Governance. *arXiv preprint arXiv:2304.04914*, 2023.

Heidari, A., Jafari Navimipour, N., Dag, H., and Unal, M. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp. e1520, 2023.

Henderson, P., Mitchell, E., Manning, C., Jurafsky, D., and Finn, C. Self-destructing models: Increasing the costs of harmful dual uses of foundation models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 287–296, 2023.

Horizon. Horizon Public Service, 2024. URL https://horizonpublicservice.org/. Accessed: 2024-06-03.

Ippolito, D., Tramèr, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette-Choo, C. A., and Carlini, N. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pp. 28–53. Association for Computational Linguistics, 2023.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. AI alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Jones, B. D. and Baumgartner, F. R. *The politics of attention: How government prioritizes problems*. University of Chicago Press, 2005.

Kaminski, M. E. Regulating the Risks of AI. *Forthcoming, Boston University Law Review*, 103, 2023.

Kitroeff, N., Gelles, D., and Nicas, J. The roots of Boeing's 737 MAX Crisis: A regulator relaxes its oversight. *The New York Times, July*, 27, 2019.

Kolt, N., Anderljung, M., Barnhart, J., Brass, A., Esvelt, K., Hadfield, G. K., Heim, L., Rodriguez, M., Sandbrink, J. B., and Woodside, T. Responsible Reporting for Frontier AI Development. *arXiv preprint arXiv:2404.02675*, 2024.

Li, B., Sun, J., and Poskitt, C. M. How Generalizable are Deepfake Detectors? An Empirical Study. *arXiv preprint arXiv:2308.04177*, 2023.

Li, S., Faure, M., and Havu, K. Liability Rules for AI-Related Harm: Law and Economics Lessons for a European Approach. *European Journal of Risk Regulation*, 13 (4):618–634, 2022.

Liao, Q. V. and Xiao, Z. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap. *arXiv preprint arXiv:2306.03100*, 2023.

Liao, T., Taori, R., Raji, I. D., and Schmidt, L. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Longpre, S., Mahari, R., Muennighoff, N., Chen, A., Perisetla, K., Brannon, W., Kabbara, J., Villa, L., and Hooker, S. The Data Provenance Project. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Longpre, S., Kapoor, S., Klyman, K., Ramaswami, A., Bommasani, R., Blili-Hamelin, B., Huang, Y., Skowron, A., Yong, Z.-X., Kotha, S., Zeng, Y., Shi, W., Yang, X., Southen, R., Robey, A., Chao, P., Yang, D., Jia, R., Kang, D., Pentland, S., Narayanan, A., Liang, P., and Henderson, P. A Safe Harbor for AI Evaluation and Red Teaming. *arXiv preprint arXiv:2403.04893*, 2024a.

Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Gero, K., Pentland, S., and Kabbara, J. Data Authenticity, Consent, and Provenance for AI are all broken: what will it take to fix them? *arXiv preprint arXiv:2404.12691*, 2024b.

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., and Perrault, R. The AI Index 2023 Annual Report. *arXiv preprint arXiv:2310.03715*, April 2023.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. The AI Index 2024 Annual Report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.

McGregor, S. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15458–15463, 2021.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Mehta, D., Mondol, N., Farahmandi, F., and Tehranipoor, M. AIME: watermarking AI models by leveraging errors. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 304–309. IEEE, 2022.

Mitchell, M. How do we know how smart AI systems are? *Science*, 381(6654):eadj5957, 2023.

Moës, N. and Ryan, F. Heavy is the Head that Wears the Crown: A risk-based tiered approach to governing General Purpose AI. *The future society*, pp. 37, 2023.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

NIST. U.S. Artificial Intelligence Safety Institute — NIST, October 2023. URL https://www.nist.gov/aisi.

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., and Raji, I. D. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. *arXiv preprint arXiv:2402.17861*, 2024.

OpenMined. Privacy-Preserving Data Science, Explained. https://blog.openmined.org/private-machine-learning-explained/, May 2020. Accessed: 2024-06-03.

Ozlati, S. and Yampolskiy, R. The formalization of AI risk management and safety standards. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., Araujo, M. M., Santos, L. L., Cruz, M. A., Oliveira, E. L., et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7 (1):15, 2023.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

Parliament, E. and Council. Artificial Intelligence Act, REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. Technical report, European Parliament and Council, May 2024. URL https://data.consilium.europa.

eu/doc/document/PE-24-2024-INIT/en/p
df. Accessed: 2024-05-05.

Pouget, H. What will the role of standards be in AI governance?, April 2023. URL https://www.adalovel
aceinstitute.org/blog/role-of-standar
ds-in-ai-governance/. Accessed: 2024-01-24.

Pouget, H. and Zuhdi, R. AI and Product Safety Standards Under the EU AI Act. *Policy Commons*, 2024.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

Raji, I. D., Xu, P., Honigsberg, C., and Ho, D. Outsider oversight: Designing a third party audit ecosystem for AI governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 557–571, 2022.

Ramanna, K. *Political standards: Corporate interest, ideology, and leadership in the shaping of accounting rules for the market economy*. University of Chicago Press, 2015.

Rawte, V., Sheth, A., and Das, A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Rigaki, M. and Garcia, S. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.

Ryan-Mosley, T. and Heikkilä, M. Three things to know about the White House's executive order on AI. *MIT Technology Review*, October 2023. URL https://ww
w.technologyreview.com/2023/10/30/10
82678/three-things-to-know-about-the
-white-houses-executive-order-on-ai/#.
Accessed: 2024-01-23.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Scherer, M. U. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. JL & Tech.*, 29:353, 2015.

Schmidt, E., Work, R. O., Bajraktari, Y., Catz, S., Horvitz, E. J., Chien, S., Jassy, A., Clyburn, M. L., Louie, G., Darby, C., et al. National security commission on artificial intelligence, 2021.

Schuett, J. Risk management in the artificial intelligence act. *European Journal of Risk Regulation*, pp. 1–19, 2023.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2022.

Shimbun, Y. Japan Govt to Establish AI Safety Institute in January, December 2023. URL https://japa
nnews.yomiuri.co.jp/politics/politic
s-government/20231221-157027/. Accessed:
2024-2-1.

Shipan, C. R. and Volden, C. The mechanisms of policy diffusion. *American journal of political science*, 52(4):
840–857, 2008.

Siegmann, C. and Anderljung, M. The Brussels effect and artificial intelligence: How EU regulation will impact the global AI market. *arXiv preprint arXiv:2208.12645*, 2022.

TechCongress. Techcongress, 2024. URL https://www.
techcongress.io/. Accessed: 2024-06-03.

The White House. Blueprint for an AI Bill of Rights. Technical report, White House, October 2022. URL https://www.whitehouse.gov/wp-conte
nt/uploads/2022/10/Blueprint-for-an-A
I-Bill-of-Rights.pdf. Accessed: 2024-01-11.

The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Technical report, The White House, October 2023a. URL https://www.federalregister.
gov/documents/2023/11/01/2023-24283/s
afe-secure-and-trustworthy-developme
nt-and-use-of-artificial-intelligenc
e. Accessed: 2024-1-11.

The White House. FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, October 2023b. URL https:
//www.whitehouse.gov/briefing-room/st
atements-releases/2023/10/30/fact-s
heet-president-biden-issues-executi
ve-order-on-safe-secure-and-trustwo
rthy-artificial-intelligence/. Accessed:
2024-01-23.

The White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, July 2023c. URL https://www.whit
ehouse.gov/briefing-room/statements-r
eleases/2023/05/04/readout-of-white-h
ouse-meeting-with-ceos-on-advancing-r
esponsible-artificial-intelligence-i
nnovation/. Accessed: 2023-6-2.

Trager, R., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L., Kreps, S., Lall, R., Larter, O., hÉigeartaigh, S. Ó., et al. International governance of civilian AI: A jurisdictional certification approach. *arXiv preprint arXiv:2308.15514*, 2023.

UN Technology Envoy. Call for nominations, August 2023. URL `https://www.un.org/techenvoy/content/call-nominations-1`. Accessed: 2024-05-27.

U.S. Bureau of Labor Statistics. Occupational Employment and Wages, May 2023. Web Page, 2023. URL `https://www.bls.gov/oes/current/oes_mat.htm`.

Vincent, J. Meta's powerful AI language model has leaked online — what happens now?, 03 2023. URL `https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse`. Accessed: 2024-01-20.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltỳnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., and Waddington, L. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):26, 2023.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., et al. Sociotechnical safety evaluation of generative AI systems. *arXiv preprint arXiv:2310.11986*, 2023.

Weissinger, L. B. AI, Complexity, and Regulation. *The Oxford Handbook of AI Governance*, February 2022.

White House OSTP. A Call to Service for AI Talent in the Federal Government. Online article, jan 2024. URL `https://www.whitehouse.gov/ostp/news-updates/2024/01/29/a-call-to-service-for-ai-talent-in-the-federal-government/`. Accessed: 27 May 2024.

Whittlestone, J. and Clark, J. Why and How Governments Should Monitor AI Development. *arXiv preprint arXiv:2108.12427*, 2021.

Wolf, Y., Wies, N., Levine, Y., and Shashua, A. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

Wu, T. In Regulating A.I., We May Be Doing Too Much. And Too Little., 11 2023. URL `https://www.nytimes.com/2023/11/07/opinion/biden-ai-regulation.html`. Accessed: 2024-01-23.

Zhang, H., Guo, Z., Zhu, H., Cao, B., Lin, L., Jia, J., Chen, J., and Wu, D. On the Safety of Open-Sourced Large Language Models: Does Alignment Really Prevent Them From Being Misused? *arXiv preprint arXiv:2310.01581*, 2023.

Zhao, X., Ananth, P., Li, L., and Wang, Y.-X. Provable robust watermarking for AI-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. Don't Make Your LLM an Evaluation Benchmark Cheater. *arXiv preprint arXiv:2311.01964*, 2023.

Zweben, S. and Bizot, B. 2016 Taulbee Survey. *The Computing Research Association*, 29(5), 2017. URL `https://cra.org/resources/taulbee-survey/`.