

**UCLA**

**UCLA Public Law & Legal Theory Series**

**Title**

Distinguishing Task-Specific and General-Purpose AI in Regulation

**Permalink**

<https://escholarship.org/uc/item/7675b1qf>

**Authors**

Selbst, Andrew

Wang, Jennifer

Venkatasubramanian, Suresh

et al.

**Publication Date**

2025-06-30

# Distinguishing Task-Specific and General-Purpose AI in Regulation

Jennifer Wang  
Stanford University  
Stanford, CA, USA  
jennjwang@stanford.edu

Solon Barocas  
Microsoft Research  
New York City, NY, USA  
solon@microsoft.com

Andrew D. Selbst  
University of California, Los Angeles  
Los Angeles, CA, USA  
aselbst@law.ucla.edu

Suresh Venkatasubramanian  
Brown University  
Providence, RI, USA  
suresh@brown.edu

## Abstract

Over the past decade, policymakers have developed a set of regulatory tools to ensure AI development aligns with key societal goals. Many of these tools were initially developed in response to concerns with task-specific AI and therefore encode certain assumptions about the nature of AI systems and the utility of certain regulatory approaches. With the advent of general-purpose AI (GPAI), however, some of these assumptions no longer hold, even as policymakers attempt to maintain a single regulatory target that covers both types of AI.

In this paper, we identify four distinct aspects of GPAI that call for meaningfully different policy responses. These are the generality and adaptability of GPAI that make it a poor regulatory target, the difficulty of designing effective evaluations, new legal concerns that change the ecosystem of stakeholders and sources of expertise, and the distributed structure of the GPAI value chain.

In light of these distinctions, policymakers will need to evaluate where the past decade of policy work remains relevant and where new policies, designed to address the unique risks posed by GPAI, are necessary. We outline three recommendations for policymakers to more effectively identify regulatory targets and leverage constraints across the broader ecosystem to govern GPAI.

## CCS Concepts

• **Social and professional topics** → **Governmental regulations.**

## Keywords

AI Governance, Regulatory Target, AI Value Chain

### ACM Reference Format:

Jennifer Wang, Andrew D. Selbst, Solon Barocas, and Suresh Venkatasubramanian. 2026. Distinguishing Task-Specific and General-Purpose AI in Regulation. In *Symposium on Computer Science and Law (CSLAW '26)*, March 3–5, 2026, Berkeley, CA, USA. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3788646.3789523>



This work is licensed under a Creative Commons Attribution 4.0 International License. *CSLAW '26*, March 3–5, 2026, Berkeley, CA, USA  
© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2447-3/26/03  
<https://doi.org/10.1145/3788646.3789523>

## 1 Introduction

Recent advances in AI have drawn intense interest from policymakers. In the past year alone, policymakers have proposed and enacted a litany of governance measures, from the European Union AI Act to a growing number of state and local initiatives in the United States [48, 62]. This wave of legislative activity has generated new confusion and contestation over how AI should be defined and governed. Some treat generative and general-purpose AI (GPAI) as a fundamentally new class of technology whose scale and autonomy distinguish it from others [4, 23], advocating for entirely new laws and institutions [28, 125]. Others caution against reinventing the regulatory wheel, instead favoring the adaptation of existing laws [46].

A notable source of this confusion is the failure to distinguish between task-specific systems and GPAI. Some regulatory efforts draw this line explicitly – for instance, the E.U. AI Act imposes distinct obligations on GPAI providers [18]. Many state and local proposals, however, do not. New York’s A6453-B, for example, defines AI as a system that “can ... make predictions, recommendations, or decisions,” yet its provisions targeting model-weight theft are clearly aimed at generative systems [123]. We also see the reverse: laws designed around task-specific systems being imposed on GPAI. New Mexico HB 60 describes AI as a technology that can be used to “generate outputs, including content, decisions, predictions, or recommendations,” yet it requires the disclosure of “sources or owners of the datasets” and the “factors in the data” used to produce outputs – assumptions that fit task-specific models with identifiable datasets and features, but not GPAI models [37].

This confusion matters because, over the past decade, policymakers have designed their regulatory toolkit around particular features of task-specific AI, but many of these features lose their relevance in a generative setting. For years before ChatGPT was released to the public, regulators have applied sector-specific frameworks to address the use of task-specific systems in employment, credit, housing, and public benefits [45, 52, 89]. These regimes were built around assumptions of how these systems are trained and evaluated for well-defined tasks, and how they are deployed to specific domains with legal guardrails. GPAI, however, breaks many of these assumptions.

In this paper, we identify four important differences between general-purpose and task-specific AI that policymakers should consider in governing both types of systems. First, the generality and adaptability of GPAI makes it a poor regulatory target: a single

model can be repurposed across many domains and use cases, so rules aimed at the model itself may fail to track where risk actually arises. Second, GPAI systems are difficult to evaluate in reliable ways because they lack task specificity. Third, GPAI raises new legal concerns that were absent or far less salient in task-specific settings, shifting the ecosystem of stakeholders and expertise required for oversight. Fourth, GPAI is developed and deployed through a highly distributed value chain which complicates the attribution of responsibility. These distinctions create governance challenges that necessitate new regulatory tools to enforce both existing and new laws.

In light of these distinctions, policymakers will need to evaluate where the past decade of policy work remains fit for purpose and where new policies are needed to address the risks unique to generative AI. Consequently, we offer three recommendations. First, that regulation should target use cases and impact, rather than the technical specification of AI systems. Second, that technical proxies should be used sparingly in regulation, with a careful balancing of risk, benefits, and a consideration of alternatives. Third, that regulation should focus on the entire *risk chain* for a use case and reinforce structural constraints that can limit harms, rather than focusing on AI systems alone.

## 2 Conceptual Distinctions

In this paper, we draw a distinction between models built to perform specific tasks, typically referred to as *task-specific*, *predictive*, or *discriminative*, and those constructed without a specific task in mind, often referred to as *general-purpose* (including those that are later fine-tuned for a specific task) or *generative*. While the distinction we raise may not capture every technical detail, it reflects the features most pertinent to policymakers.<sup>1</sup>

### 2.1 Task-specific AI systems

Task-specific AI systems are designed to *predict* a particular outcome and are almost exclusively built using supervised learning. Building such a system starts with a *task* described by training data consisting of input vectors and the “ground truth” predictions. The goal is to build a system that can correctly reproduce the ground truth predictions given the inputs, in the hope that it will accurately predict outcomes for a new input.

While this description is an oversimplification of model training, it captures the most salient aspects of task-specific AI for policymakers—there is an intended task, the task can be expressed in terms of specific training data, and the goal is to build a system that performs this task. [150].

### 2.2 General-purpose AI systems

When we turn to GPAI, the picture is much more complicated. Some policy definitions draw distinctions from task-specific systems based on the nature of the output—content versus predictions [134]. This is reflected in AI definitions in legislative proposals

that insert the word ‘content’ into the list of outputs that might be produced by an AI system [87, 144, 160].

Others focus on the development process and nature of the model. For example, Bommasani et al. [21] describe it as a model built using *self-supervised learning* and that has billions of parameters. Lee et al. [98] rejects the term “foundation model” in favor of “generative AI”, which they describe as characterized by a mixture of models, built at scale, that makes use of specific components like transformers.

The EU AI Act defines “general purpose AI” as “an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications” [62]. This definition retains the method of production (self-supervision), the reference to scale, and the implied capabilities.

In line with these policy discussions, we view a GPAI system as one that is built using self-supervision, that usually (though not necessarily) is described using billions of parameters, and that most importantly is not trained to perform any specific task beyond that which the self-supervision requires.<sup>2</sup> Self-supervision is perhaps the most distinctive feature of what characterizes GPAI models. In language models, for example, self-supervision proceeds by training the language model to fill in the blanks in lines of text in which words have been masked randomly.<sup>3</sup> The rationale for this form of training is that for a model to be able to correctly predict missing words in arbitrary sentences in text, it must be able to build an internal representation of meaning in text that will allow it to complete the sentences. Regardless of whether models are able to represent ‘meaning’, it is the case that models do end up constructing an internal representation space for words (or tokens) and use this internal representation to respond to prompts or other natural language queries. This internal representation is what allows GPAI systems to perform a wide range of downstream tasks for which it might not have ever been trained specifically.<sup>4</sup>

Most of the other elements used to define GPAI systems—the architecture of transformers, the apparent need for billions of parameters, and so on—are either consequences of the training process or reflections of prevailing design practices. These features represent our current understanding of how to build models effectively, but they are not necessary to capture the task-agnostic nature of GPAI systems.

But what of the myriad further steps in training a GPAI model, like fine-tuning, reinforcement learning from human feedback, inference-tuning, and so on? First, most of these post-pretraining steps do not fundamentally change the fact that a GPAI model relies on an internal representation to perform a wide range of downstream tasks, even if they might customize the model to perform

<sup>1</sup>The EU AI Act, among others, makes this same distinction, but calls out the former systems as ‘AI’ and the latter as ‘general-purpose AI’. Bommasani et al. [21] refer to the former as ‘task-specific’ (and the latter as ‘task-agnostic’, but settle on ‘foundation’). We feel that the terms ‘task-specific’ and ‘general-purpose’ are useful for the distinctions that we wish to explore, but note that there are several different ways to refer to these systems, and there is no clear consensus.

<sup>2</sup>To illustrate the difficulty even with this rendition, systems like BERT [53] are built using self-supervision, but are not generative in the sense of producing new, fluent sequences of text.

<sup>3</sup>The model is trained using supervised learning with a specific task—predict the missing words—but using the training data itself as inputs and targets for training. This gives rise to the term “self-supervision.” For image models, the process is different, but is still self-supervised.

<sup>4</sup>While some self-supervised systems are built for specific tasks, policy discussions focus on GPAI systems where self-supervision enables broad, task-agnostic capabilities.

more effectively on certain tasks. Second, these downstream training steps apply to an already built model, rather than creating a new one from scratch. Unlike task-specific models, where using the model for a new task usually involves rebuilding it from scratch, fine-tuning merely modifies an existing (and functioning) model, and can be performed without knowledge of how the original model was built.<sup>5</sup> In turn, the builders of the base model need not have any knowledge of how fine-tuning might occur and for what purpose. A GPAI model that is fine-tuned to perform a prediction task is still a GPAI model for the purpose of the points we discuss in this paper.

### 3 What makes general-purpose AI distinctive?

In this section, we outline four key distinctions that set GPAI apart—its general-purpose design, challenges in evaluation, emergence of new legal concerns, and distributed value chain structure. These differences challenge the assumptions underlying existing regulatory instruments, revealing how oversight and accountability mechanisms designed for task-specific systems lose their utility when applied to GPAI systems.

#### 3.1 General-purpose AI systems make poor regulatory targets because of their generality and adaptability.

GPAI systems are general-purpose and adaptable, able to handle unstructured inputs across many modalities (e.g., text, images, audio, etc.). As a result, the space of potential applications for GPAI is vast. Through fine-tuning and prompting, the same base model can be adapted to perform remarkably different tasks without specific training. For example, GPT-4 can understand and generate human-like text across various topics and styles, process and generate images, engage in real-time verbal conversations, perform complex reasoning tasks, and write and debug code, all without fundamentally changing the underlying model [133].

In contrast, task-specific systems are purpose-built to perform a defined task. They streamline, support, and sometimes fully automate human decision-making processes. Although these systems vary in their technical implementation, they share a common function as tools for decision-making. This shared function gives rise to a shared set of risks across different applications. Whether applied to hiring, credit scoring, or tenant screening, task-specific AI raises a similar set of concerns related to fairness, transparency, and accountability. This creates a coherent basis for regulators to establish overarching guardrails that target universal harms posed by task-specific AI—one that draws on principles of due process, non-discrimination, and privacy. For example, the Blueprint for an AI Bill of Rights outlines common expectations for automated decision systems, such as the right to notice and explanation and protection from algorithmic bias, regardless of the domain [180]. Similarly, the proposed AI Civil Rights Act seeks to prohibit the use, sale, or promotion of algorithmic decision-making systems that discriminate on the basis of protected characteristics [108]. The comprehensive AI regulation passed in Colorado in 2024 similarly

sets requirements for developers of "high-risk" AI systems across a wide range of domains [44].

The generality and adaptability of GPAI breaks down this regulatory coherence. GPAI systems can be put to a far more diverse set of uses and although these applications may stem from the same technical foundation, many will raise distinct risk profiles. Like task-specific AI, GPAI can be used for decision-making, including in the high-stakes context of hiring [66], lending [24], and housing [104], thus invoking the same set of regulatory concerns. But it can also be used to generate children's stories, which raises concerns over age-appropriate material and creative rights [168, 171], for drafting legal documents, where there may be concerns over hallucinations and the confidentiality risks of client information [161], and for coding assistance, which could introduce security vulnerabilities and lower code quality [72, 139]. While this generality and adaptability is not inherently problematic, as there are sector-specific frameworks that govern each of these use cases, the challenge arises when treating GPAI as the regulatory target (i.e., the specific entity that regulation is designed to govern).

Because there is no shared regulatory objective (i.e., the specific goal or intended outcome that regulation aims to achieve) across generative use cases, efforts to develop regulations at the model level may be misguided. Model-level regulations that apply across diverse applications would require abstracting away all of the details that actually account for the concerns raised by these specific applications. Because GPAI systems lack a shared purpose, the only features shared across applications tend to be technical properties of the models, such as compute usage and model size, which do not correspond to any particular applications of concern. As a result, when regulators attempt to pursue broad upstream controls specifically targeting GPAI, they are left with few footholds other than these technical properties. These features are attractive from a regulatory perspective because they are easy to define and measure. However, they are often poor proxies for the many harms that could arise from different applications.

Thus, regulation targeting GPAI systems often yields rules that are broad in scope, but shallow in substance. These rules are *shallow* because they only address the highest-level factors that could affect the risks posed by GPAI—typically only those observable at the model level without consideration of the deployment context. This creates a regulatory framework that appears expansive but lacks the specificity to ensure meaningful oversight. We elaborate on the limitations of regulating at the model level in 4.1, where we instead recommend regulating the applications and impacts of these systems.

#### 3.2 General-purpose AI systems are difficult to evaluate due to their lack of task specificity.

Evaluation of AI systems has become a touchstone for oversight. In many cases, evaluations are designed to assess system performance using quantitative metrics, such as accuracy and recall, that typically require a known ground truth for comparison. By systematically assessing AI system performance, evaluations provide enforcement mechanisms and remedies to contest false claims and ineffective systems [120, 142]. They are crucial to determining whether a system is fit for purpose and sufficiently reliable for deployment. In other cases, evaluations are designed to assess models' impacts.

<sup>5</sup>The knowledgeable reader will note that transfer learning is exactly the process of taking a given model and adapting it using small amounts of new data for a related task. While this is true, it is typically less common in predictive settings to employ transfer learning since it is more straightforward to rebuild the model.

One common example is the use of evaluation to identify when task-specific AI systems produce disparate impacts—that is, when they differentially benefit or burden different groups. From the E.U. AI Act to New York City local laws, policymakers have called for independent audits, impact assessments, and risk assessments in order to understand both basic functionality (i.e., performance) and harms (i.e., impacts), including discrimination, privacy, and physical safety [62, 122, 181]. The widespread adoption of audit requirements and impact assessments in U.S. congressional bills [173], state-level initiatives [31, 43], and international proposals [62, 84, 85] reflects their importance in providing assurance and accountability in AI systems.

Task-specific systems are designed around tasks that map from specific input spaces to output spaces (see Section 2). This means developers are able to conduct evaluations within bounded problem domains where decision-making objectives can be clearly defined and measured. Typically, evaluators use a holdout set—data withheld from training that contains known outcomes that serve as the ground truth for assessing the quality of the models’ predictions [55]. By comparing the model’s outputs to these outcomes, evaluators can compute a set of metrics (e.g., accuracy) to measure how well the system performs on a predetermined task. The evaluation process naturally mirrors training by repurposing the same data pipelines and directly measuring the model’s performance on the task it was trained on [96].

The task-specificity of these systems has three important implications for evaluations. First, evaluations of task-specific systems have **self-evident targets**. Because the model is trained with a clear objective, evaluators can measure how well the system performs its intended function. Second, evaluations can be **properly scoped**. That is, evaluators can define the input space for each task-specific system and bound its functionality to specific use cases. This allows evaluators to assess model performance, knowing that the model will only operate within this defined input space and purpose. Third, evaluations can be **reproducible**. When provided with the same inputs, outputs, and metrics, different evaluators can independently reproduce and verify the system’s performance. These three properties allow policymakers to establish certain expectations around the validity and reliability of evaluations. Although there remain technical and institutional challenges in evaluations for task-specific systems, the expectations are that (i) evaluations should appropriately characterize the expected behavior of the systems relative to its defined objective (targeted), (ii) they should achieve sufficient coverage of the system’s intended uses (well-scoped), and (iii) the evaluation results should be consistent and repeatable across different independent evaluators (reproducible).

The absence of task specificity in GPAI systems means that these properties fall away. GPAI systems are not trained for specific tasks but rather with broader objectives to learn structural patterns in the data (e.g. predicting the next word given the previous words). As such, there is no clear target against which to evaluate their performance [178]. Evaluators instead target general capabilities or properties, such as linguistic understanding [179], reasoning [40, 41], and biases [14, 137, 159]. These capabilities or properties are characterized by the model’s responses to a set of custom-built and annotated scenarios, typically in the form of multiple-choice

or binary questions [58, 59, 79, 179]. A scenario is meant to instantiate a desired use case for GPAI models, and a collection of scenarios form benchmarks that guide and measure progress in AI development [59].

While capabilities-based evaluations are meant to produce evidence of likely performance on a broad range of possible tasks that would presumably draw on relevant capabilities, the relationship between capabilities and performance on specific, real-world tasks is often quite tenuous. As many have observed, the particular tasks used in capabilities-focused benchmarks don’t often resemble the tasks that GPAI models are used to perform in practice [25, 49, 60, 110, 112, 141, 145, 156].

Even when tasks are well-specified, the real-world performance and impact of GPAI models are shaped by how those tasks are carried out via fine-tuning, prompting, and other forms of downstream adaptations [25, 39, 103, 141]. These post-training modifications are often highly variable and unpredictable [13], leading the same model to exhibit markedly different behavior even when used to perform what seem to be the same or similar tasks. Given how sensitive evaluation results can be to these downstream choices, it can be exceedingly difficult to reliably predict performance based on evaluations that only consider a limited set of possible choices.

Evaluation at the foundation model level falls prey to what Selbst et al. call the “framing trap”—the failure to account for the broader sociotechnical system in which AI operates [150]. With GPAI, the framing trap is largely unavoidable if one seeks to evaluate the foundation model. The sociotechnical context of the AI system varies with each use case, shaped by the distinct patterns of user interactions, institutional norms, and task-specific objectives [15]. As a result, even if it is feasible to develop robust, standardized evaluation methods for the foundation model itself, evaluating models in isolation from the context of its uses cannot adequately surface concrete harms or support meaningful mechanisms of accountability, which would be seen or apply only to the *combination* of model and application context.

While past regulatory requirements around AI audits and impact assessments have hinged on the assumption that developers can meaningfully evaluate AI systems prior to deployment, GPAI challenges the reliability of current evaluation methods as regulatory tools. One example that highlights its limitations is California’s S.B. 1047. The vetoed state bill would have required model developers “to assess whether the covered model is reasonably capable of causing or materially enabling critical harm” before making it publicly available [30]. The problem, however, is that evaluating harm at the model level, without accounting for downstream use cases, fails to consider the role of user interactions, implementation decisions, and the organizational processes in shaping the real-world impacts of model use. As such, regulatory proposals hitched to pre-deployment evaluations of GPAI models may misdiagnose where and how harms actually materialize, leading to ineffective or misguided oversight.

### 3.3 General-purpose AI opens up new areas of legal concern that were not implicated by task-specific AI.

Recent developments make clear that GPAI poses legal challenges largely absent from task-specific AI. As such, the responses of existing legal regimes will necessarily differ, each requiring distinct domain-specific expertise. Copyright offers the most prominent example to the general public. While task-specific AI can raise copyright concerns at the input stage—whether training on protected works constitutes infringement [102, 153]—this was not the primary legal concern with the technology. GPAI not only magnifies concerns about infringement in training, but also implicates the reproduction right due to its ability to output text, images, and videos, which task-specific AI did not [78, 98, 100]. AI companies currently face major infringement lawsuits based on using protected content in model training and producing outputs that closely resemble copyrighted works [7, 36, 68, 71].

Beyond copyright concerns, GPAI has drawn scrutiny for its role in spreading defamation and misinformation. GPAI systems regularly fabricate plausible-sounding falsehoods, producing content with invented names, fake quotations, and non-existent sources [133]. Already, lawsuits have been filed over alleged libel created by GPAI programs [107, 175]. As GPAI systems are increasingly integrated into search engines and virtual assistants, the risk of reputational harm and public deception intensifies. In more extreme cases, GPAI systems have been used to commit fraud. Synthetic voice technologies have been used to clone speech patterns of real individuals, enabling impersonation through robocalls [9, 64]. Hyperrealistic deepfake videos and AI-generated images have similarly been used to deceive viewers and distort perception, and pose particular difficulties for evidentiary processes in courts [16, 126]. The proliferation of synthetic content raises the question of whether and when AI developers should be held liable for content generated by its systems [98, 177].

While task-specific systems have long played a role in intelligence analytics and surveillance, GPAI introduces new concerns in chemical, biological, radiological, and nuclear (CBRN) security [117, 167]. Traditionally, the development of CBRN weapons has required high levels of expertise, access to restricted materials, and coordination across domains. However, GPAI systems could lower informational barriers and expand the capabilities of malicious actors by synthesizing vast bodies of literature and proposing novel ideas for CBRN weapon development [69, 114, 154]. While the actual realization of such threats still depends on real-world capabilities [138], the possibility that GPAI models could support CBRN threats has raised concerns about their dual-use potential, driving new discussions about risk mitigation, early warning systems, and access controls [11, 116, 132, 162].

The legal challenges posed by GPAI in new domains highlight the limitations of treating AI as a singular regulatory target. These issues fall under distinct legal regimes, each with its own doctrines, precedents, and norms that have developed over decades. For example, copyright law has long grappled with questions of authorship, originality, and fair use, and has developed specific doctrinal tests to differentiate between protected creative works and derivative copies [1]. CBRN security governance has been grounded in the

precautionary principle, the practice of taking preventative action in the face of uncertainty. It has relied on nonproliferation regimes, material control measures, and international oversight to restrict the development and spread of high-risk threats [169]. Each of these domains has also drawn on distinct communities of experts and stakeholders with different priorities, risk tolerances, and terminologies. The differences in how each community conceptualizes and addresses the challenges posed by GPAI mean that regulatory approaches that work well in one domain may be ineffective or counterproductive in another. As a result, effective governance for GPAI will require broader and more diverse domain knowledge than is required for task-specific AI alone. Attempts to consolidate these distinct legal challenges into a single, overarching AI policy framework risks not only oversimplifying the regulatory landscape but also overlooking the stakeholders and experts best positioned to guide policy decisions.

### 3.4 General-purpose AI is built on a highly distributed value chain that complicates accountability.

AI systems are developed and deployed through complex, interdependent networks of data providers, compute providers, model developers, hosting services, model adapters, and application developers [22, 95, 98]. Collectively, these actors comprise the "AI value chain," a sequence of modular processes that "contribute towards the production, deployment, use, and functionality of AI technologies" [42]. This modularity allows for specialization and efficient division of labor [140], but it also disperses control and obscures lines of responsibility [26, 35, 182].

Task-specific AI systems typically exhibit a tightly integrated value chain, from data curation to feature selection to model development. The system's single purpose and structure are understood at the time of creation, and the creation itself is often done collaboratively between developer and deployer [50]. Thus, the universe of potential harms is either known, or at least foreseeable, to both the developer and deployer of the system.

Generative AI complicates this picture. As model development is no longer tied to specific tasks, providers focus on building large, incredibly complex, general-purpose models that are designed to be customized and reused. Then, in the minimally complex case, a deployer will fine-tune the system for their use case and then deploy it. There are several important implications to this change for legal concepts related to responsibility and accountability.

The first is that the conceptual distance between the model developer's work and the ultimate harm is much larger. With a task-specific model, developers will be aware of the final use, and will often work in concert with their customers; now, with a GPAI model, they are building general-purpose technologies, and many ultimate uses will be hard, if not impossible, to foresee [182]. Additionally, the complexity of these models renders specific outcomes likely unforeseeable, even where the general domain of use can be predicted in advance [42]. Foreseeability, in turn, is a bedrock limitation to legal liability in most instances [74]. This translates to a general aversion to holding creators of general-purpose systems with substantial lawful uses liable for downstream harms [155]. Telephone

creators are not held liable for crimes organized over the phone, nor are computer makers for harms researched and planned on laptops.

The second is that there are more—and more specialized—roles in the overall value chain. This is not inherently difficult for a legal regime to handle. For example, the EU’s AI Act lays out different sets of responsibilities for providers, deployers, importers, and distributors [62]. But new classes of intermediate actors—model adapters and integrators—are getting involved. Model adapters fine-tune or modify GPAI models for specific tasks or domains, while integrators incorporate models into applications or systems, shaping how users interact with and experience them. The fluidity of the different value chains and the disaggregation of responsibilities into different roles can make it difficult for a statute to stay current [56].

This fragmentation of roles is not inherently problematic for a legal regime. In product liability, for instance, all parties in the chain of commerce are held liable for an injury to a plaintiff, and the court lets them sort out who ultimately pays. In the EU AI Act, a deployer who makes substantial modifications or changes the purpose of a deployed system becomes a provider as a legal matter [62, Article 25]. As long as *someone* is liable in the end, then oversight can be accomplished. What’s different here is that lines between actors in the GPAI value chain are often blurry, making the specific questions of how responsibility is divided more difficult and raising the cost of oversight [184]. Sticking with products liability as an example, under the so-called “component parts doctrine,” a supplier of a component in a larger product is liable for the harm only if the component itself is defective, the integration causes the component to be defective, or the maker of the component substantially participates in the integration [129]. This doctrine relies on a continued conceptual separation between the component and the larger product after integration. What’s different now is that the downstream uses actually *morph* the product at issue. Once another actor fine-tunes a model, adapts it to a new context, or integrates it into a larger application, the resulting system cannot be cleanly disaggregated back into its original components to determine who is responsible for which aspects of its behavior. This lack of separability means that courts will need to examine in detail how the system was modified, what changes those modifications introduced, and how those changes are connected to the harm at issue. Untangling this highly convoluted chain of modifications and actors will raise both the duration and expense of litigation.

Third, the extra roles and distributed global nature of the AI value chain may pose jurisdictional difficulties in legal accountability. Models can be developed in one country and be ported across jurisdictions trivially. Yet the legality of their training or deployment may differ by jurisdiction. For example, if personal data of any EU residents was used to train GPAI models, the training likely violated the GDPR, so what should EU regulators make of someone who independently fine-tunes one of the existing models for a local purpose? Indeed, Yew et al. [185] argue that companies are likely to exploit the geographical scoping in the AI Act to continue to provide services to customers in the EU while avoiding governance requirements in the Act. Current US-EU disputes over whether the Digital Services Act should apply to US-based LLM providers, and its implication for free speech protections within the United States are yet another example of the cross-border jurisdictional challenges that GPAI presents. These challenges are amplified in the

case of open-source or open-weight models, which are published and freely accessible anywhere. In contrast, these kinds of concerns are less common with task-specific models, as they are typically bespoke systems trained on proprietary data and developed with infrastructure confined to a single jurisdiction.

None of this means that upstream providers should be exempt from responsibility for downstream harms. For one, they maintain access to core system components, and retain the ability to control the systems they release. Because there are only a small number of upstream providers, they have immense market power [166]. Further, GPAI has an unprecedented demand for compute and data. While earlier task-specific systems might be trained on thousands or millions of examples, GPAI models rely on billions or trillions of data points [2]. On average, training large-scale GPAI models requires roughly 100 times more compute than earlier systems [151]. As a result, upstream providers have become systematically important within the GPAI value chain, as downstream actors increasingly rely on a few providers for access to pretrained models, large-scale data, and advanced compute infrastructure—resources foundational to building and scaling GPAI systems [42]. These intensified dependencies give upstream providers an outsized influence over who can build, scale, and commercialize GPAI, a dynamic that was far less pronounced in the context of task-specific AI.

## 4 Recommendations

Whether emerging technology drives policy has been a longstanding question in technology law, including AI governance [33, 47, 90]. Drawing on past lessons from tech policymaking, we propose three recommendations for policymakers to consider when it comes to GPAI. First, regulators should identify regulatory targets based on the *function* and *consequences* of AI systems rather than technical implementation. Second, only in *narrow circumstances* should the technical properties of AI systems be employed to determine the level of regulatory scrutiny. Third, regulators should examine the entire *risk chain* of AI applications and strengthen structural constraints that could serve as more effective barriers to misuse.

### 4.1 Regulators should generally identify regulatory targets based on the applications and consequences of AI systems rather than technical implementation.

The challenge of regulating AI exemplifies many of the core sources of contention and difficulties in tech policy. Central to this discourse is the distinction between tech-neutral and tech-specific regulatory approaches. Tech-neutral regulation governs activities and consequences, while tech-specific regulation governs specific technical implementations.

The case for tech-neutral regulation has taken on new significance in the context of AI. In task-specific AI, model development is oriented to specific use cases: this task specificity facilitates compliance with tech-neutral regulation, as we argue in Section 3.1. However, for GPAI, this specificity is lost due to the separation between development and deployment. Instead, the general-purpose abilities of GPAI have led to calls for regulation at the level of the model itself, leading to shallow approaches based on model parameters and other technical proxies. Bommasani [19] has identified

two types of proxies commonly used in AI policy: (1) resource expenditure (i.e., the amount of compute in FLOPs and the amount of data in modality-specific units), and (2) model properties (i.e. capabilities benchmarks and safety benchmarks).

As the AI landscape evolves, however, rigid regulatory boundaries defined by technical proxies such as compute thresholds or model size quickly lose relevance. While these metrics offer an initial assessment of the model’s capabilities, they are limited. Technical proxies represent static snapshots that may neither reflect changes in model capabilities over time, nor behaviors that arise from user interactions in real-world scenarios. In fact, we already see systems that achieve state-of-the-art capabilities with significantly reduced compute needs and dramatic performance gains from simple modifications in adaptation procedures [51], rendering such proxies ineffective as regulatory boundaries [38, 176, 186]. By directing regulatory attention to the deployment context and intended use, regulators can better calibrate the level of regulatory scrutiny to the application-specific risks while accommodating ongoing technical progress.

AI regulation that relies on technical proxies risks not just becoming outdated but being both under- and over-inclusive [47].<sup>6</sup> Regulation is under-inclusive when it leaves out systems that achieve similar concerning outcomes through alternative methods. For instance, relying on compute thresholds to identify frontier AI models may inadvertently exclude systems that cause harm through less resource-intensive methods (e.g., misinformation bots) [124]. Regulation is over-inclusive when it captures more systems than intended. For example, the same compute thresholds could sweep in benign applications like climate modeling that happen to share high computational demands. The challenge in defining the regulatory scope is compounded by the broad range of contexts in which the same AI systems can be deployed to serve vastly different purposes [73]. The generality and adaptability of GPAI therefore underscore the need for regulatory frameworks that account for the specific contexts and outcomes of their use [115].

Tech-neutral regulation also helps address the challenge of regulatory arbitrage, where actors might switch to alternative technical approaches to circumvent tech-specific rules while achieving the same concerning outcomes [29, 47]. This is especially relevant given the rapid pace of AI development, where new architectures and methodologies emerge frequently. Tying regulations to specific technologies creates incentives for companies to design their systems in ways that fall outside the regulatory scope [185].

One example that highlights the limits of tech-specific regulation is the European Union’s legislation on genetically modified organisms (GMO) [136]. The legislation targets crops that have been altered by recombinant DNA technology to address public concerns over potential health risks. In response, seed producers have turned

away from recombinant DNA technology toward mutagenesis, using mutagenic chemicals or radiation treatments to induce random genetic changes in plants [93]. While these methods are not considered recombinant DNA technology, they achieve the same outcome of producing genetically modified crops with commercially desirable traits. Ironically, mutagenic crops may pose comparable or even greater risks than GMO crops on food safety and the environment [94, 113]. This case demonstrates how tech-specific regulation can incentivize the development of what Dan Burk calls “perverse innovations” that “ingeniously dodge the intended outcome of regulation, while still formally adhering to the text of the regulation” [29].

## 4.2 Identifying regulatory targets by technical specifications can still make sense as a last resort.

In recent years, the AI regulatory landscape has seen a chorus of concerns over AI safety and security risks [65, 105, 164, 165]. These concerns have led to a surge in tech-specific proposals that aim to regulate AI systems, with model properties as technical proxies for their capacity to precipitate harm in society [30, 62, 162, 172].

Despite its limitations, tech-specific regulation has garnered support in AI oversight due to its feasibility, political expediency, and capacity for preemptive intervention [47, 130]. Compute governance, for instance, establishes concrete thresholds, the amount of computing resources used to train a model, which can be easily verified and applied across different actors and systems [149]. This establishes objective, measurable intervention points in AI development, providing clarity as to whether and how regulation applies to an actor or artifact and increasing consistency in the application of the law. Tech-specific approaches also enable regulators to precisely target oversight toward certain types of technology while minimizing the burden on others. By promulgating separate rules for AI models with different computational requirements and specifications, regulators can establish clear standards for compliance and enforcement without considering the breadth and depth of AI applications, each with its own legal implications and risk profile. Finally, tech-specific regulation allows for intervention before potentially harmful systems are developed or deployed. Regulation through technology can be appropriate where the goal is to make something impossible rather than merely illegal [101, 143]. This is particularly important for AI systems that could pose catastrophic risks or irreversible damage, where relying on ex post remedies may be inadequate [12].

However, as we highlighted in the prior section, tech-specific regulation is not without its limitations. Its application should be carefully constrained to avoid unintended consequences, such as overregulating benign use cases and creating regulatory loopholes. For this reason, we recommend that technical proxies should be employed for regulation only in narrow circumstances where three conditions are met: (1) tech-neutral regulation is demonstrably impractical in addressing the risk of concern, (2) the risk being addressed is severe, plausible, and attributable to the existence and configuration of the technology, and (3) the benefits from tech-specific regulation outweigh the potential harms. We now expand on each of these three conditions.

First, tech-specific regulation should only be used when tech-neutral regulation is impossible or impractical. A common critique

<sup>6</sup>Considered more precisely, this is a point about legal rules and standards, not tech neutrality [47, p. 406]. Where regulations are more rule-like, they can be over- and under-inclusive. We make the point here about tech-specificity because we think it is likely that tech-specificity leads to rules more often than standards. This is both because the tendency for tech-specific regulation is already motivated by a desire for easier administrability and because technical standards and specifications are often stated in quantitative terms. Examples of tech-specific standard-based regulation may become more prevalent if governments, for example, mandate that LLMs have output filters that accomplish some goal like fairness or a more complicated set of goals in the vein of Anthropic’s attempt at “Constitutional AI” [8], but for now, we think the assumption holds.

of tech-neutral regulation is that it may be ineffective and even impossible to implement, by the time harms of a technology become apparent through widespread adoption [67]. For instance, in nuclear safety regulation, the long-term environmental and health impacts from radioactive contamination and nuclear meltdowns are so severe and irreversible that a tech-neutral approach is inadequate [174]. Instead, nuclear safety regulation relies on highly specific rules tailored to the unique risks of nuclear technology, such as strict access to nuclear materials and safety protocols that govern nuclear facilities [183]. Similarly, GPAI presents a severe consequence that task-specific AI does not: risk of catastrophic harm [80]. Catastrophic harm refers to events with potentially devastating consequences for human society, such as the misuse of AI to create bioweapons or the loss of control over advanced AI systems [157]. Unlike more predictable risks, catastrophic harms are difficult to quantify because they often involve scenarios with little to no historical data, making it nearly impossible to assign a precise probability of occurrence [63]. Some caution that the race among companies to advance AI heightens the risk that these models could pose unforeseen threats, calling for pauses in AI development [65]. The concern that GPAI might be used for a broad range of dangerous and difficult-to-anticipate applications necessitates tech-specific regulation to preemptively mitigate these risks rather than reacting after the fact, when it may be too late to avert disaster.

Second, tech-specific regulation should be reserved for cases where the potential harm is plausible and evidence-based, significant, and attributable to the specific technology or technological configuration. An example is AI-generated nonconsensual intimate imagery (NCII) based in the likeness of real people, otherwise known as deepfakes. AI-generated NCII is a current problem, with significant and well-documented harm [81, 97, 111]. While NCII is starting to be addressed with criminalization, civil liability, and/or mandated takedowns [170], NCII may also be a candidate for tech-specific regulation because the harm is accomplished immediately, cheaply, and at scale, making it harder to detect and remove [86, 163]. An example of such tech-specific regulation would be a requirement that image generators have filters that prevent the generation of NCII with features of real people [117]. This is tech-specific because it regulates through a configuration of the technology, rather than addressing the harm directly, like liability for its creation or takedown requirements do [32, 109]. So far, the technological limitations have not been totally successful—ability to evade them being one of the risks of tech-specific regulation—but something could be created that performs better, and successfully prevents the harm [54].

Third, tech-specific regulation may be justified when the risks posed by a specific technology are significant enough that the benefits of regulation interventions outweigh the costs. One example of this could be facial recognition technology (FRT). Hartzog and Selinger have argued that FRT’s mere existence is uniquely dangerous in a way that justifies banning the technology outright, due in part to the unique role of the face in our identity and social interactions [75]. In fact, some state and local governments have already enacted bans on its use [152]. These restrictions curb the indiscriminate deployment of facial recognition technology by state actors, to prevent mass privacy violations, wrongful arrests, and disproportionate targeting of marginalized communities. While facial

recognition technology may offer potential benefits in security and efficiency, ban advocates argue that these advantages are grossly overshadowed by the severe and evidenced risk it poses [76, 82]. GPAI may present analogous concerns in biosecurity. It risks facilitating a bio-engineered pandemic by (1) lowering the barrier to entry for non-experts to develop biological weapons, and (2) discovering novel, more lethal pathogens that evade current safeguards. The former may reduce the cost and expertise required to create harmful biological agents, while the latter raises the ceiling of harm by increasing the lethality and transmissibility of pathogens. These two outcomes establish distinct threat models and introduce different trade-offs that regulators will need to evaluate and prioritize when weighing the benefits of AI advancements for bio research against the costs of potential misuse.

Regardless of how narrowly regulatory targets are defined, regulators must maintain the flexibility to update the basis for legal scrutiny [19]. Even proponents of compute governance acknowledge that regulatory thresholds—such as a training compute threshold of  $10^{26}$  operations—or list-based controls on technologies can quickly become outdated as technology evolves [149]. This necessitates establishing mechanisms, from the outset, to periodically re-evaluate and revise policies. Importantly, this does not mean updating a single threshold, but rather ensuring frameworks evolve in response to new developments and evidence [34]. To provide this adaptability in AI governance, Bommasani et al. have called for forward-looking blueprints that map societal conditions (e.g., specific AI capabilities) to appropriate policy responses [20].

#### 4.3 Regulators should focus on structural constraints and dependencies that prevent harm.

The current discourse on AI regulation has predominantly focused narrowly on the AI value chain. While this has yielded concrete technical governance proposals [27, 132, 149], it neglects the systemic nature of AI risks and impacts, as its harm manifests through interactions across an even broader ecosystem. We propose an evaluation framework to assess regulatory approaches across domains, using biosecurity as a case study. The framework examines where existing regulations assume conditions that GPAI now disrupt, and pinpoints where new measures are needed to close new gaps.

One way to conceptualize the interactions between actors, systems, and resources across an ecosystem is through a risk chain: a sequence of actions that traces how potential harms are conceived, developed, and realized across an ecosystem. In biosecurity, a risk chain might begin with an actor having an “irresponsible, misguided, or malicious” intention [147]. This intention gives rise to a biological idea, which is then converted into biological data—such as a pathogen genome—and transformed into a live biological artifact. The artifact is subsequently cultured, tested, and ultimately dispersed into the target environment [147]. In the following, we use bioweapons development as a case study to study how harm progresses or is prevented from one stage to another.

Scholars have long observed that the particular arrangements and affordances of digital technology create extra-legal structures and contours that shape conduct in practice, often in ways that regulation and laws adapt to and build on [101, 131, 143, 158]. These

structural “constraints” function as “regulators of behavior that prevent conduct through technological or physical barriers in the world” [158]. These constraints can be physical (e.g., access to infrastructure), economic (e.g., material costs), or epistemic (e.g., domain expertise).<sup>7</sup> Much like law, these constraints can make certain activities prohibitively difficult or expensive to carry out at scale. They shape which harms can propagate along a risk chain and which are effectively contained by the ecosystem.

Bioweapons development offer a clear example of how structural constraints operate within a risk chain. Even if an actor begins with a harmful biological idea, progressing through the chain requires access to biomaterials and specialized lab infrastructure [10, 138]. This requirement imposes physical constraints for obtaining and handling hazardous biomaterials and highlights risk chain dependencies, as the movement and use of biomaterials necessarily involve interactions with institutions that produce and distribute these agents [6]. These dependencies make certain transitions in the risk chain, such as transforming a digital genome into a cultured pathogen, contingent on navigating the physical, economic, and epistemic barriers. At the same time, these points offer opportunities for policy interventions, including screening and monitoring the distribution of biomaterials [5, 92], requiring licenses for researchers working with high-risk biological agents [91], and enforcing strict security protocols in laboratory facilities [119].

We propose a three-step domain-independent framework to help identify and evaluate how AI interacts with the risk chain. **First, regulators should identify the key dependencies and constraints at each stage of the risk chain.** These could include both deliberately engineered dependencies that make certain actions infeasible without authorization, as well as by-product constraints of the technological and physical environment raising the difficulty, cost, or time required to carry out harmful activities [158]. In biosecurity, multiple constraints and dependencies impede the progression of the risk chain, making it more difficult to carry out biological attacks. To begin, there are epistemic barriers due to a lack of biological knowledge and scientific training [118, 135]. Next, actors need biomaterials, specialized equipment, state-of-the-art methods and deep domain expertise [99]. And the final stages of culturing, testing, and deploying biological agents are time- and resource-intensive [61, 118, 128].

**Second, regulators should assess the extent to which existing regulations depend on structural constraints to remain effective.** When these constraints shift or weaken due to technological changes, the regulatory regimes built around them may lose their force. In the case of biosecurity, regulatory efforts to deter the creation of bioweapons have traditionally relied on physical constraints, such as controlling access to hazardous biomaterials and enforcing safety protocols within high-containment laboratory facilities [6, 127, 169]. These efforts rest on the assumption that malicious actors cannot operationalize biological threats without overcoming significant physical barriers.

**Third, regulators should evaluate how advancing AI capabilities could undermine or bypass the structural constraints that existing regimes rely on.** Structural constraints lose their regulatory force when actors can leverage AI to circumvent the friction and barriers they impose at different stages of the risk chain. This could occur in three key ways. First, AI could collapse stages of the risk chain by streamlining multiple-step processes to a single, more efficient operation. In the biorisk chain, early stages of literature review, hypothesis formulation and experiment design can be bypassed entirely by LLMs fine-tuned on biological data [3, 10, 121, 146]. Second, AI could discover new ways to formulate bioweapons using readily accessible materials, rather than known inputs that are subject to access controls [10, 77, 106, 148, 154], thereby compromising biosecurity regimes that rely on physical containment. Third, AI could reduce friction by accelerating workflows, automating processes, and lowering the time and cost that would otherwise slow down or discourage certain activities [17, 83, 146, 154]. These three mechanisms demonstrate how AI can reshape the structure and pace of the risk chain, potentially undermining constraints that once served as effective regulatory barriers.

The extent to which these constraints remain effective depends on the regulatory objectives and assumptions underlying the structure of the risk chain. For instance, even if GPAI lowers the barrier to entry to biodevelopment—by expanding access to biological information and reducing the time and cost needed for pathogen development—the physical constraints likely remain intact [70, 88]. Alternatively, if regulation focused on epistemic constraints, such as restricting access to sensitive methodological knowledge by redacting details in published biology manuscripts [57], then the regulatory emphasis would shift to limiting knowledge diffusion. Regulators should determine which constraints are most critical to preserving the function of existing legal regimes and assess how AI may undercut their regulatory force.

## 5 Conclusion

Our goal in this work is to illustrate when the differences between general-purpose and task-specific AI are merely matters of degrees—where existing regulatory approaches remain applicable—and when they are differences in kind, demanding new regulatory instruments and forms of intervention. We identify four distinct areas of meaningful difference and make three recommendations for policymakers based on these differences. We believe that these recommendations can help policymakers avoid many of the fault lines in current AI policy debates and lead to more effective policymaking for GPAI.

<sup>7</sup>We note that these categories of barriers are not fixed or mutually exclusive. In practice, they often overlap and reinforce one another. For instance, a lack of technical knowledge (epistemic) may require hiring experts (economic), which in turn may be necessary to operate specialized systems (physical). While analytically distinct, these barriers often operate together as part of a broader regulatory environment.

## References

- [1] 17 U.S.C. § 102(b). n.d.. 17 U.S.C. § 102(b). <https://www.law.cornell.edu/uscode/text/17/102>. Accessed: 2025-05-05.
- [2] AI Now Institute. 2023. AI Now 2023 Landscape Report: Confronting Tech Power. <https://ainowinstitute.org/wp-content/uploads/2023/04/AI-Now-2023-Landscape-Report-FINAL.pdf> Accessed: 2025-05-23.
- [3] Gregory C. Allen. 2023. Advanced Technology: Examining Threats to National Security. Testimony before the U.S. Senate Committee on Homeland Security and Governmental Affairs. <https://www.hsgac.senate.gov/wp-content/uploads/Allen-Testimony.pdf> Accessed: 2025-05-21.
- [4] Sam Altman, Greg Brockman, and Ilya Sutskever. 2023. Governance of Superintelligence. <https://openai.com/blog/governance-of-superintelligence>. Accessed: 2025-05-07.
- [5] HHS ASPR. 2010. Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA. <https://aspr.hhs.gov/legal/syndna/Documents/syndna-guidance.pdf> Accessed: 2025-05-21.
- [6] HHS ASPR. 2025. Biosafety Levels and Biorisk Management. <https://aspr.hhs.gov>. Accessed: 2025-05-21.
- [7] Authors Guild et al. v. OpenAI Inc. et al. 2023. Authors Guild et al. v. OpenAI Inc. et al. <https://www.courtlistener.com/docket/67810584/authors-guild-v-openai-inc/> Case No. 1:23-cv-08292-SHS, U.S. District Court for the Southern District of New York.
- [8] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] <https://arxiv.org/abs/2212.08073>
- [9] Sam Barrington, Emily A. Cooper, and Hany Farid. 2025. People are poorly equipped to detect AI-powered voice clones. 11004 pages. <https://doi.org/10.1038/s41598-025-94170-3>
- [10] Steph Batalis. 2023. AI and BioRisk: An Explainer. <https://cset.georgetown.edu/publication/ai-and-biorisk-an-explainer/> Accessed: 2025-05-21.
- [11] Steph Batalis. 2024. Anticipating Biological Risk. <https://doi.org/10.51593/20240013> Accessed: 2025-05-05.
- [12] Yoshua Bengio. 2023. Written Testimony of Professor Yoshua Bengio: Presented before the U.S. Senate Judiciary Subcommittee on Privacy, Technology, and the Law. [https://yoshuabengio.org/wp-content/uploads/2023/07/Written-Testimony-and-biography-of-Yoshua-Bengio\\_U.S.-Senate-Judiciary-Subcommittee-on-Privacy-Technology-and-the-Law\\_25\\_07\\_2023.pdf](https://yoshuabengio.org/wp-content/uploads/2023/07/Written-Testimony-and-biography-of-Yoshua-Bengio_U.S.-Senate-Judiciary-Subcommittee-on-Privacy-Technology-and-the-Law_25_07_2023.pdf)
- [13] Jan Betley, Daniel Tan, Niels Warnecke, Anna Szyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. 2025. Emergent Misalignment: Narrow finetuning can produce poorly misaligned LLMs. arXiv:2502.17424 [cs.CL] <https://arxiv.org/abs/2502.17424>
- [14] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv:2110.01963 [cs.CV] <https://arxiv.org/abs/2110.01963>
- [15] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. 5454–5476 pages. <https://doi.org/10.18653/v1/2020.acl-main.485>
- [16] Matyas Bohacek and Hany Farid. 2025. The making of an AI news anchor—and its implications. e2315678121 pages. <https://doi.org/10.1073/pnas.2315678121>
- [17] Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. arXiv:2304.05332 [physics.chem-ph] <https://arxiv.org/abs/2304.05332>
- [18] Claire Boine and David Rolnick. 2023. Why the AI Act Fails to Understand Generative AI. <https://doi.org/10.2139/ssrn.4644701> Presented at We Robot 2023, Boston University School of Law; Winner of the Best Paper Award.
- [19] Rishi Bommasani. 2023. Drawing Lines: Tiers for Foundation Models. <https://crfm.stanford.edu/2023/11/18/tiers.html>
- [20] Rishi Bommasani, Sanjeev Arora, Yejin Choi, Daniel E. Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, Fei-Fei Li, Arvind Narayanan, Alondra Nelson, Emma Pierson, Joelle Pineau, Gaël Varoquaux, Suresh Venkatasubramanian, Ion Stoica, Percy Liang, and Dawn Song. 2024. A Path for Science- and Evidence-based AI Policy. Understanding AI Safety. <https://understanding-ai-safety.org/>
- [21] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models.
- [22] Rishi Bommasani, Dilara Soylu, Thomas I. Liao, Kathleen A. Creel, and Percy Liang. 2023. Ecosystem Graphs: The Social Footprint of Foundation Models. arXiv:2303.15772 [cs.LG] <https://arxiv.org/abs/2303.15772>
- [23] Nick Bostrom. 2014. Superintelligence: Paths, Dangers, Strategies. OCLC: ocn881706835.
- [24] Donald E. Bowen III, S. McKay Price, Luke C. D. Stein, and Ke Yang. 2024. Measuring and Mitigating Racial Disparities in Large Language Model Mortgage Underwriting. <https://doi.org/10.2139/ssrn.4812158>
- [25] Samuel R. Bowman and George Dahl. 2021. What Will it Take to Fix Benchmarking in Natural Language Understanding? 4843–4855 pages. <https://doi.org/10.18653/v1/2021.naacl-main.385>
- [26] Ian Brown. 2023. Expert Explainer: Allocating Accountability in AI Supply Chains. <https://www.adalovelaceinstitute.org/resource/ai-supply-chains/>. Retrieved October 15, 2023.
- [27] Ryan Browne. 2024. Tech giants pledge AI safety commitments — including a ‘kill switch’ if they can’t mitigate risks. <https://www.cnn.com/2024/05/21/tech-giants-pledge-ai-safety-commitments-including-a-kill-switch.html> Accessed: 2025-05-21.
- [28] Bureau of Industry and Security, U.S. Department of Commerce. 2025. Framework for Artificial Intelligence Diffusion. *Federal Register*, Vol. 90, No. 10, pp. 4544–4579. <https://www.federalregister.gov/documents/2025/01/15/2025-00636/framework-for-artificial-intelligence-diffusion> Federal Register Document Number: 2025-00636. Accessed: 2025-05-07.
- [29] Dan L. Burk. 2016. Perverse Innovation. <https://scholarship.law.wm.edu/wmlr/vol58/iss1/2>
- [30] California Senate. 2024. Senate Bill 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act. <https://legiscan.com/CA/text/SB1047/id/2999979> Accessed: 2025-05-07.
- [31] California State Legislature. 2019. Senate Bill No. 36: Pretrial release: risk assessment tools. [https://leginfo.ca.gov/faces/billTextClient.xhtml?bill\\_id=20190200SB36](https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=20190200SB36) Accessed: 2025-05-01.
- [32] California State Legislature. 2024. Senate Bill No. 926: Crimes—Distribution of Intimate Images. California 2023–2024 Regular Session. <https://legiscan.com/CA/bill/SB926/2023> Chaptered by Secretary of State, Chapter 289, Statutes of 2024. Introduced by Senators Wahab and Becker.
- [33] Ryan Calo. 2015. Robotics and the Lessons of Cyberlaw. 513–563 pages. <https://digitalcommons.law.uw.edu/faculty-articles/23>
- [34] Jeremy Carp. 2018. Autonomous Vehicles: Problems and Principles for Future Regulation. <https://ssrn.com/abstract=3305990> Available at SSRN: <https://ssrn.com/abstract=3305990>
- [35] Sarah Huiyi Cen, Aspen Hopkins, Andrew Ilyas, Aleksander Madry, Isabella Struckman, and Luis Videgaray Caso. 2023. AI Supply Chains. <https://ssrn.com/abstract=4789403>. <https://doi.org/10.2139/ssrn.4789403> Available at SSRN.
- [36] Chabon et al. v. OpenAI, Inc. 2023. Chabon et al. v. OpenAI, Inc. <https://www.courtlistener.com/docket/67778017/chabon-v-openai-inc/> Case No. 3:23-cv-04625, U.S. District Court for the Northern District of California.
- [37] Christine Chandler, Andrea Romero, Debra M. Sariñana, Linda M. Trujillo, and Heather Berghmans. 2025. NM House Bill 60. Legislative document, First Session of the 57th Legislature.
- [38] Caowei Chen. 2025. How a top Chinese AI model overcame US sanctions. <https://www.technologyreview.com/2025/01/24/1110526/china-deepseek-top-ai-despite-sanctions/> Accessed: 2025-05-07.
- [39] Wayne Chi, Valerie Chen, Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, Aditya Mittal, Naman Jain, Tianjun Zhang, Ion Stoica, Chris Donahue, and Ameet Talwalkar. 2025. Copilot Arena: A Platform for Code LLM Evaluation in the Wild. arXiv:2502.09328 [cs.SE] <https://arxiv.org/abs/2502.09328>
- [40] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. arXiv:2202.04053 [cs.CV] <https://arxiv.org/abs/2202.04053>
- [41] François Chollet. 2019. On the Measure of Intelligence. arXiv:1911.01547 [cs.AI] <https://arxiv.org/abs/1911.01547>
- [42] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding Accountability in Algorithmic Supply Chains. 1186–1197 pages. <https://doi.org/10.1145/3593013.3594073>
- [43] Colorado General Assembly. 2021. Senate Bill 21-169: Protecting Consumers from Unfair Discrimination in Insurance Practices. <https://leg.colorado.gov/bills/sb21-169> Accessed: 2025-05-01.
- [44] Colorado General Assembly. 2024. Senate Bill 24-205: Concerning consumer protections in interactions with artificial intelligence systems. <https://leg.colorado.gov/bills/sb24-205> Accessed: 2025-09-27.
- [45] Consumer Financial Protection Bureau. 2022. Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms. <https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms>. [Accessed 23-12-2024].
- [46] Jack Corrigan and Owen J. Daniels. 2024. Don’t Reinvent the Wheel to Govern AI. Net Politics, Council on Foreign Relations. <https://www.cfr.org/blog/dont-reinvent-wheel-govern-ai> Accessed 2025-05-09.
- [47] Rebecca Crotoof and BJ Ard. 2021. Structuring Techlaw. 347– pages. <https://doi.org/10.2139/ssrn.3664124> Univ. of Wisconsin Legal Studies Research Paper

- No. 1761.
- [48] Heather Curry. 2024. 2024 State Summary on AI. <https://techpost.bsa.org/2024/10/22/2024-state-summary-on-ai/>
  - [49] Ernest Davis. 2023. Benchmarks for Automated Commonsense Reasoning: A Survey. arXiv:2302.04752 [cs.AI] <https://arxiv.org/abs/2302.04752>
  - [50] Daswin De Silva and Daminda Alahakoon. 2022. An artificial intelligence life cycle: From conception to production. 100489 pages. <https://doi.org/10.1016/j.patter.2022.100489>
  - [51] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] <https://arxiv.org/abs/2501.12948>
  - [52] Department of Housing and Urban Development. 2024. Guidance on Application of the Fair Housing Act to the Screening of Applicants for Rental Housing. [https://archives.hud.gov/news/2024/FHEO\\_Guidance\\_on\\_Screening\\_of\\_Applicants\\_for\\_Rental\\_Housing.pdf](https://archives.hud.gov/news/2024/FHEO_Guidance_on_Screening_of_Applicants_for_Rental_Housing.pdf). Provides guidance on nondiscriminatory use of tenant screening technologies, including AI and machine learning, under the Fair Housing Act.
  - [53] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 4171–4186 pages.
  - [54] Michelle L. Ding and Harini Suresh. 2025. The Malicious Technical Ecosystem: Exposing Limitations in Technical Governance of AI-Generated Non-Consensual Intimate Images of Adults. arXiv:2504.17663 [cs.HC] <https://arxiv.org/abs/2504.17663>
  - [55] David Donoho. 2017. 50 Years of Data Science. 745–766 pages. <https://doi.org/10.1080/10618600.2017.1384734> Accessed: 2025-05-01.
  - [56] Alex C. Engler and Andrea Renda. 2022. *Reconciling the AI Value Chain with the EU's Artificial Intelligence Act*. Technical Report. Centre for European Policy Studies (CEPS). <https://www.ceps.eu/cepspublications/reconciling-the-ai-value-chain-with-the-eus-artificial> Retrieved October 15, 2023.
  - [57] Martin Enserink. 2011. Grudgingly, Virologists Agree to Redact Details in Sensitive Flu Papers. <https://www.science.org/content/article/grudgingly-virologists-agree-redact-details-sensitive-flu-papers> Accessed: 2025-05-21.
  - [58] Aarohi Srivastava et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615 [cs.CL] <https://arxiv.org/abs/2206.04615>
  - [59] Percy Liang et al. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL] <https://arxiv.org/abs/2211.09110>
  - [60] Kavin Ethayarajah and Dan Jurafsky. 2021. Utility is in the Eye of the User: A Critique of NLP Leaderboards. arXiv:2009.13888 [cs.CL] <https://arxiv.org/abs/2009.13888>
  - [61] European Commission. 2005. Synthetic Biology: Applying Engineering to Biology. Report of a NEST High-Level Expert Group. [https://haseloff.plantsci.cam.ac.uk/resources/SynBio\\_reports/NEST\\_syntheticbiology\\_2005.pdf](https://haseloff.plantsci.cam.ac.uk/resources/SynBio_reports/NEST_syntheticbiology_2005.pdf) Structuring the European Research Area — Anticipating Scientific and Technological Needs.
  - [62] European Parliament. 2024. Regulation (EU) 2024/1689 of the European Parliament: Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). <http://data.europa.eu/eli/reg/2024/1689/oj/eng>
  - [63] Daniel A. Farber. 2011. Uncertainty. 901–957 pages. <https://doi.org/10.2139/ssrn.1555343> UC Berkeley Public Law Research Paper No. 1555343.
  - [64] Federal Communications Commission. 2024. FCC Makes AI-Generated Voices in Robocalls Illegal. <https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal> Declaratory Ruling FCC 24-17.
  - [65] Future of Life Institute. 2023. Pause Giant AI Experiments: An Open Letter. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> Accessed 2025-05-09.
  - [66] Johann D. Gaebler, Sharad Goel, Aziz Huq, and Prasanna Tambe. 2025. Auditing large language models for race & gender disparities: Implications for artificial intelligence-based hiring. 23794607251320229 pages. <https://doi.org/10.1177/23794607251320229>
  - [67] David Gee. 2009. Late Lessons from Early Warnings: Towards Realism and Precaution with EMF? 217–231 pages. <https://doi.org/10.1016/j.pathophys.2009.01.004> Epub 2009 May 21, PMID: 19467848.
  - [68] Getty Images (US), Inc. v. Stability AI, Inc. 2023. Getty Images (US), Inc. v. Stability AI, Inc. <https://docs.justia.com/cases/federal/district-courts/delaware/decede/1:2023cv00135/81407/1> Case No. 1:23-cv-00135, U.S. District Court for the District of Delaware.
  - [69] Anjali Gopal, Nathan Helm-Burger, Lennart Justen, Emily H. Soice, Tiffany Tzeng, Geetha Jeyapragasan, Simon Grimm, Benjamin Mueller, and Kevin M. Esvelt. 2023. Will releasing the weights of future large language models grant widespread access to pandemic agents? arXiv:2310.18233 [cs.AI] <https://arxiv.org/abs/2310.18233>
  - [70] Daniel Grushkin, Todd Kuiken, and Piers Millett. 2013. Seven Myths & Realities About Do-It-Yourself Biology. [https://www.wilsoncenter.org/sites/default/files/media/documents/publication/7\\_myths\\_final.pdf](https://www.wilsoncenter.org/sites/default/files/media/documents/publication/7_myths_final.pdf)
  - [71] Michael M. Grynbaum and Ryan Mac. 2023. New York Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html> Accessed: 2025-05-04.
  - [72] Bill Harding. 2025. GitClear AI Code Quality Research Pre-Release. [https://www.gitclear.com/blog/gitclear\\_ai\\_code\\_quality\\_research\\_pre\\_release](https://www.gitclear.com/blog/gitclear_ai_code_quality_research_pre_release) Accessed: 2025-05-23.
  - [73] Laurie Harris. 2021. *Artificial Intelligence: Background, Selected Issues, and Policy Considerations*. Technical Report R46795. Congressional Research Service. <https://crsreports.congress.gov/product/pdf/R/R46795> Accessed: 2025-05-07.
  - [74] H.L.A. Hart and Tony Honore. 1985. Causation in the Law. 255 pages.
  - [75] Woodrow Hartzog and Evan Selinger. 2018. Facial Recognition Is the Perfect Tool for Oppression. Medium. <https://medium.com/@hartzog/facial-recognition-is-the-perfect-tool-for-oppression-df2b6e50e1e3> Co-authored with Evan Selinger.
  - [76] Drew Harwell and Eva Dou. 2020. Huawei Tested AI Software That Could Recognize Uighur Minorities and Alert Police, Report Says. <https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/> Accessed May 9, 2025.
  - [77] Helena. 2023. Biosecurity in the Age of AI. <https://www.helenabiosecurity.org/> Accessed: 2025-05-21.
  - [78] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. *Foundation Models and Fair Use*. Technical Report. Stanford Law and Economics Olin Working Paper No. 584. <https://ssrn.com/abstract=4404340> Available at SSRN: <https://ssrn.com/abstract=4404340> or <http://dx.doi.org/10.2139/ssrn.4404340>.
  - [79] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY] <https://arxiv.org/abs/2009.03300>
  - [80] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks. arXiv:2306.12001 [cs.CY] <https://arxiv.org/abs/2306.12001>
  - [81] Nicola Henry, Clare McGlynn, Asher Flynn, Kelly Johnson, Anastasia Powell, and Asher J. Scott. 2020. Image-based Sexual Abuse: A Study on the Causes and Consequences of Non-consensual Nude or Sexual Imagery. <https://doi.org/10.4324/9781351135153>
  - [82] Kashmir Hill. 2020. Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match. <https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html> Updated January 6, 2021. Accessed May 9, 2025.
  - [83] Takashi Inagaki, Akari Kato, Koichi Takahashi, Haruka Ozaki, and Genki N. Kanda. 2023. LLMs can generate robotic scripts from goal-oriented instructions in biological laboratory automation. arXiv:2304.10267 [q-bio.QM] <https://arxiv.org/abs/2304.10267>
  - [84] Information Commissioner's Office. 2023. Annex A: Fairness in the AI Life-cycle. Guidance on the AI Auditing Framework Draft Guidance for Consultation. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/> Accessed: 2025-05-01.
  - [85] Innovation, Science and Economic Development Canada. 2023. Artificial Intelligence and Data Act (AIDA). <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act> Accessed: 2025-05-01.
  - [86] Internet Watch Foundation. 2024. AI CSAM Report Update: What has changed in the AI CSAM landscape? Internet Watch Foundation. <https://www.iwf.org.uk/media/nadcb1z/iwf-ai-csam-report-update-public-jul24v13.pdf> Update to the October 2023 AI report.
  - [87] Iowa State Legislature. 2025. Iowa House Bill HF 406. <https://www.legis.iowa.gov/legislation/BillBook?ga=91&ba=HF406> 91st General Assembly.
  - [88] Catherine Jefferson, Filippa Lentzos, and Claire Marris. 2014. Synthetic Biology and Biosecurity: Challenging the “Myths”. 115 pages. <https://doi.org/10.3389/fpubh.2014.00115>
  - [89] Elisa Jillson. 2021. Aiming for Truth, Fairness, and Equity in Your Company's Use of AI. <https://privacysecurityacademy.com/wp-content/uploads/2021/04/Aiming-for-truth-fairness-and-equity-in-your-companys-use-of-AI.pdf>. Federal Trade Commission.
  - [90] Meg Leta Jones. 2018. Does Technology Drive Law? The Dilemma of Technological Exceptionalism in Cyberlaw. <https://doi.org/10.2139/ssrn.2981855>
  - [91] C.B. Jonsson, K.S. Cole, C.J. Roy, D.S. Perlin, G. Byrne, and members of the RBL-NBL Directors Network. 2013. Challenges and Practices in Building and Implementing Biosafety and Biosecurity Programs to Enable Basic and Translational Research with Select Agents. 12634 pages. <https://doi.org/10.4172/2157-2526.S3-015>
  - [92] Alexander Kane and Michael T. Parker. 2024. Screening State of Play: The Biosecurity Practices of Synthetic DNA Providers. 85–95 pages. <https://doi.org/10.1089/apb.2023.0027>
  - [93] Jack Kaskey. 2013. Mutant Crops Drive BASF Sales Where Monsanto Denied: Commodities. Published Nov. 13, 2013, 11:04 AM. Available at: <http://www.bloomberg.com/news/articles/2013-11-13/mutant-crops-drive-basf-sales-where-monsanto-denied-commodities>. Archived at: <https://perma.cc/57JJ-BHD7>.

- [94] Jack Kaskey. 2013. The Scariest Veggies of Them All. *Bloomberg*, Nov. 22, 2013, 5:52 PM. Available at: <http://www.bloomberg.com/news/articles/2013-11-21/monsanto-vs-dot-mutant-crop-developers-in-global-seed-market>. Archived at: <https://perma.cc/S7Y6-KM2T>.
- [95] Thalia Khan and Madhulika Srikumar. 2024. From Code to Consumer: PAI's Value Chain Analysis Illuminates Generative AI's Key Players. <https://partnershiponai.org/from-code-to-consumer-pais-value-chain-analysis-illuminates-generative-ais-key-players/>.
- [96] Ron Kohavi. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1137–1143 pages. <https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf>
- [97] Elizabeth Laird, Maddy Dwyer, and Kristin Woelfel. 2024. In Deep Trouble: Surfacing Tech-Powered Sexual Harassment in K-12 Schools. Center for Democracy & Technology. <https://cdt.org/insights/in-deep-trouble-surfacing-tech-powered-sexual-harassment-in-k-12-schools/> Report by the Center for Democracy & Technology.
- [98] Katherine Lee, A. Feder Cooper, and James Grimmelmman. 2024. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain. arXiv:2309.08133 [cs.CY] <https://arxiv.org/abs/2309.08133>
- [99] Milton Leitenberg, Raymond A. Zilinskas, and Jens H. Kuhn. 2012. The Soviet Biological Weapons Program: A History. <https://www.jstor.org/stable/j.ctt2jbscf>
- [100] Mark A. Lemley. 2023. How Generative AI Turns Copyright Upside Down. <https://ssrn.com/abstract=4517702> Available at SSRN: <https://ssrn.com/abstract=4517702> or <http://dx.doi.org/10.2139/ssrn.4517702>.
- [101] Lawrence Lessig. 2006. Code: Version 2.0. <https://lessig.org/product/codev2/> Edited by Benkler, Yochai. Available under a Creative Commons license..
- [102] Amanda Levendowski. 2018. How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem. 579 pages. Available at SSRN: <https://ssrn.com/abstract=3024938>.
- [103] Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap.
- [104] Eric Justin Liu, Wonyoung So, Peko Hosoi, and Catherine D'Ignazio. 2024. Racial Steering by Large Language Models: A Prospective Audit of GPT-4 on Housing Recommendations. <https://doi.org/10.1145/3689904.3694709>
- [105] Garrison Lovely. 2024. Laws Need to Catch Up to Artificial Intelligence's Unique Risks. <https://www.nytimes.com/2024/09/29/opinion/ai-risks-safety-whistleblower.html> Guest Essay, Opinion Section.
- [106] David Luckey, Sara Duhachek Muggy, Taylor Frey, David Stebbins, Tracey Rissman, Bianca Espinosa, Daniel Tapia, Greg McKelvey Jr., Neeti Pokhriyal, Joseph Dawson, et al. 2025. *Mitigating Risks at the Intersection of Artificial Intelligence and Chemical and Biological Weapons*. Technical Report RR-A2990-1. RAND Corporation. [https://www.rand.org/pubs/research\\_reports/RR-A2990-1.html](https://www.rand.org/pubs/research_reports/RR-A2990-1.html) Accessed: 2025-05-21.
- [107] Mark Walters v. OpenAI. 2024. <https://law.justia.com/cases/federal/appellate-courts/ca11/23-13843/23-13843-2024-04-01.html> United States Court of Appeals for the Eleventh Circuit.
- [108] Sen. Edward J. Markey. 2024. S.5152 - Artificial Intelligence Civil Rights Act of 2024. <https://www.congress.gov/bills/118th/congress/senate-bill/5152>. 118th Congress, Introduced September 24, 2024. Referred to the Senate Committee on Commerce, Science, and Transportation. Accessed: 2025-05-23.
- [109] Massachusetts General Court. 2024. H.4241 - An Act to Prevent Abuse and Exploitation. 193rd General Court (2023–2024). <https://malegislature.gov/Bills/193/H4241> Sponsored by the House Committee on Ways and Means. Passed House and Senate with amendments; under conference as of June 2024.
- [110] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. arXiv:1902.01007 [cs.CL] <https://arxiv.org/abs/1902.01007>
- [111] Clare McGlynn, Erika Rackley, and Ruth Houghton. 2017. Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse. <https://doi.org/10.1007/s10691-017-9343-2>
- [112] Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence. arXiv:2402.09880 [cs.AI] <https://arxiv.org/abs/2402.09880>
- [113] Shane H. Morris. 2007. EU Biotech Crop Regulations and Environmental Risk: A Case of the Emperor's New Clothes? *Trends in Biotechnology* 25, 1 (January 2007), 2–6. <https://doi.org/10.1016/j.tibtech.2006.11.004> Epub 2006 Nov 20.
- [114] Christopher A. Mouton, Caleb Lucas, and Ella Guest. 2024. *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*. Technical Report. RAND Corporation. [https://www.rand.org/pubs/research\\_reports/RR-A2977-2.html](https://www.rand.org/pubs/research_reports/RR-A2977-2.html)
- [115] Dewey Murdick, James Dunham, and Jennifer Melot. 2020. *AI Definitions Affect Policymaking*. Technical Report. Center for Security and Emerging Technology. <https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Definitions-Affect-Policymaking.pdf> Accessed: 2025-05-07.
- [116] National Academies of Sciences, Engineering, and Medicine. 2025. The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations. <https://doi.org/10.17226/28868> Accessed: 2025-05-05.
- [117] National Institute of Standards and Technology. 2024. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. Technical Report NIST AI 600-1. U.S. Department of Commerce, National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.600-1>
- [118] National Research Council. 2004. Biotechnology Research in an Age of Terrorism. <http://www.nap.edu/catalog/10827>
- [119] National Research Council. 2009. Responsible Research with Biological Select Agents and Toxins. <https://www.ncbi.nlm.nih.gov/books/NBK44957/>
- [120] National Telecommunications and Information Administration. 2024. Purpose of Evaluation. <https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/ai-system-evaluations/purpose-of-evaluation#ftn179> Accessed: 2025-05-01.
- [121] Cassidy Nelson and Sophie Rose. 2023. Understanding AI-Facilitated Biological Weapon Development. <https://doi.org/10.7117/nm7j-qzt1>
- [122] New York City Council. 2020. Int 1894-2020: A Local Law to amend the administrative code of the City of New York, in relation to requiring an annual report on automated decision systems used by city agencies. <https://legistar.council.nyc.gov/LegislationDetail.aspx?GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&ID=4344524> Accessed: 2025-05-01.
- [123] New York State Assembly. 2025. NY State Assembly Bill 2025-A6453B. [https://www.nysenate.gov/legislation/bills/2025/A6453/amendment/B Responsible AI Safety and Education \(RAISE\) Act, amending the General Business Law relating to training and use of artificial intelligence frontier models; accessed January 12, 2026.](https://www.nysenate.gov/legislation/bills/2025/A6453/amendment/B%20Responsible%20AI%20Safety%20and%20Education%20(RAISE)%20Act,%20amending%20the%20General%20Business%20Law%20relating%20to%20training%20and%20use%20of%20artificial%20intelligence%20frontier%20models)
- [124] Gavin Newsom. 2024. Veto Message for California Senate Bill 1047. Office of the Governor, State of California. <https://legiscan.com/CA/text/SB1047/id/2999979>
- [125] Michelle Nichols. 2023. UN Chief Backs Idea of Global AI Watchdog Like Nuclear Agency. <https://www.reuters.com/technology/un-chief-backs-idea-global-ai-watchdog-like-nuclear-agency-2023-06-12/> Accessed 2025-05-09.
- [126] Sophie J. Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. e2120481119 pages. <https://doi.org/10.1073/pnas.2120481119>
- [127] NSTC. 2022. *Evidence-Based Laboratory Biorisk Management: Science and Technology Roadmap*. Technical Report. Health Security Threats Subcommittee. [https://www.whitehouse.gov/wp-content/uploads/2022/04/04-2022-NSTC-ST-Biorisk-Research-Roadmap\\_FINAL.pdf](https://www.whitehouse.gov/wp-content/uploads/2022/04/04-2022-NSTC-ST-Biorisk-Research-Roadmap_FINAL.pdf)
- [128] Nuclear Threat Initiative. 2023. The Convergence of Artificial Intelligence and the Life Sciences. <https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/> Accessed: 2025-05-21.
- [129] Restatement (Third) of Torts: Product Liability. 1998. § 5.
- [130] Paul Ohm. 2010. The Argument against Technology-Neutral Surveillance Laws. 1685–1714 pages.
- [131] Paul Ohm and Jonathan Frankle. 2018. Desirable Inefficiency. 777–823 pages. <https://scholarship.law.ufl.edu/flr/vol70/iss4/2>
- [132] OpenAI. 2024. Building an Early Warning System for LLM-Aided Biological Threat Creation. <https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/> Accessed: 2025-05-04.
- [133] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] <https://arxiv.org/abs/2303.08774>
- [134] Organisation for Economic Co-operation and Development (OECD). 2024. Generative AI. <https://www.oecd.org/en/topics/sub-issues/generative-ai.html> Policy sub-issue overview.
- [135] Sonia Ben Ouagrham-Gormley. 2014. Barriers to Bioweapons. <https://www.cornellpress.cornell.edu/book/9780801452888/barriers-to-bioweapons/#bookTabs=4>
- [136] The European Parliament and the Council of the European Union. 2001. Directive 2001/18/EC of the European Parliament and of the Council of 12 March 2001 on the deliberate release into the environment of genetically modified organisms and repealing Council Directive 90/220/EEC - Commission Declaration. <http://data.europa.eu/eli/dir/2001/18/oj>
- [137] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. arXiv:2110.08193 [cs.CL] <https://arxiv.org/abs/2110.08193>
- [138] Aidan Peppin, Anka Reuel, Stephen Casper, Elliot Jones, Andrew Strait, Usman Anwar, Anurag Agrawal, Sayash Kapoor, Sanmi Koyejo, Marie Pellat, Rishi Bommasani, Nick Frosst, and Sara Hooker. 2025. The Reality of AI and Biorisk. arXiv:2412.01946 [cs.AI] <https://arxiv.org/abs/2412.01946>
- [139] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. 2023. Do Users Write More Insecure Code with AI Assistants? 2785–2799 pages. <https://doi.org/10.1145/3576915.3623157>
- [140] Michael E. Porter. 1985. Competitive Advantage: Creating and Sustaining Superior Performance.
- [141] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark.
- [142] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew D. Selbst. 2022. The Fallacy of AI Functionality. <https://doi.org/10.1145/3531146>

- 3533158 Accessed: 2025-05-01.
- [143] Joel R. Reidenberg. 1998. *Lex Informatica: The Formulation of Information Policy Rules through Technology*. 553–593 pages. [https://ir.lawnet.fordham.edu/faculty\\_scholarship/42](https://ir.lawnet.fordham.edu/faculty_scholarship/42) Originally published in 1997-1998.
  - [144] Rhode Island State Legislature. 2025. Rhode Island Senate Bill S 0627. <https://webserver.rilegislature.gov/BillText/BillText25/SenateText25/S0627.pdf> Senate Bill, 2025 Session.
  - [145] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. 4902–4912 pages. <https://doi.org/10.18653/v1/2020.acl-main.442>
  - [146] Sophie Rose, Richard Moulange, James Smith, and Cassidy Nelson. 2024. The Near-Term Impact of AI on Biological Misuse. <https://www.longtermresilience.org/wp-content/uploads/2024/07/CLTR-Report-The-near-term-impact-of-AI-on-biological-misuse-July-2024-1.pdf> Accessed: 2025-05-21.
  - [147] Anders Sandberg and Cassidy Nelson. 2020. Who Should We Fear More: Biohackers, Disgruntled Postdocs, or Bad Governments? A Simple Risk Chain Model of Biorisk. 155–163 pages. <https://doi.org/10.1089/hs.2019.0115> Epub 2020 Jun 10.
  - [148] Jonas B. Sandbrink. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. arXiv:2306.13952 [cs.CY] <https://arxiv.org/abs/2306.13952>
  - [149] Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O’Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. 2024. Computing Power and the Governance of Artificial Intelligence. arXiv:2402.08797 [cs.CY] <https://arxiv.org/abs/2402.08797>
  - [150] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. 59–68 pages. <https://doi.org/10.1145/3287560.3287598>
  - [151] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute Trends Across Three Eras of Machine Learning. <https://doi.org/10.1109/TETCI.2022.3193973>
  - [152] Nathan Sheard and Adam Schwartz. 2022. The Movement to Ban Government Use of Face Recognition. Deeplinks Blog, Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2022/05/movement-ban-government-use-face-recognition?language=fa> Accessed 2025-05-09.
  - [153] Benjamin Sobel. 2017. Artificial Intelligence’s Fair Use Crisis. 45 pages. Available at SSRN: <https://ssrn.com/abstract=3032076>.
  - [154] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. Can Large Language Models Democratize Access to Dual-Use Biotechnology? arXiv:2306.03809 [cs.CY] <https://arxiv.org/abs/2306.03809>
  - [155] Sony Corp. of America v. Universal City Studios, Inc. 1984. 417 pages. <https://supreme.justia.com/cases/federal/us/464/417/> Also known as the “Betamax case”.
  - [156] Arjun Subramonian, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. 2023. It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance. arXiv:2305.09022 [cs.CL] <https://arxiv.org/abs/2305.09022>
  - [157] Cass R. Sunstein. 2007. The Catastrophic Harm Precautionary Principle. <https://doi.org/10.2202/1539-8323.1091>
  - [158] Harry Surden. 2007. Structural Rights in Privacy. 1605–1629 pages. Available at SSRN: <https://ssrn.com/abstract=1004675>.
  - [159] Zeerak Talat, Aurélie Névél, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. 26–41 pages. <https://doi.org/10.18653/v1/2022.bigscience-1.3>
  - [160] Texas State Legislature. 2025. Texas House Bill HB 1495. <https://capitol.texas.gov/tlodocs/89R/billtext/html/HB001491.htm> 89th Legislature, Regular Session.
  - [161] The State Bar of California. 2023. Practical Guidance for the Use of Generative Artificial Intelligence in the Practice of Law. <https://www.calbar.ca.gov/Portals/0/documents/ethics/Generative-AI-Practical-Guidance.pdf>.
  - [162] The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/> Accessed: 2025-05-01.
  - [163] David Thiel et al. 2023. Generative ML and CSAM: Implications and Mitigations. Stanford Digital Repository. <https://purl.stanford.edu/jv206yg3793>
  - [164] Meaghan Tobin. 2024. A.I. Pioneers Call for Protections Against ‘Catastrophic Risks’. <https://www.nytimes.com/2024/09/16/business/china-ai-safety.html> Accessed: 2025-05-07.
  - [165] Zeynep Tufekci. 2025. The Dangerous A.I. Nonsense That Trump and Biden Fell For. <https://www.nytimes.com/2025/02/05/opinion/ai-deepseek-trump-biden.html>
  - [166] UK Competition and Markets Authority. 2023. AI Foundation Models: Initial Report. <https://www.gov.uk/government/publications/ai-foundation-models-initial-report> Published 18 September 2023.
  - [167] UK Government. 2024. New Commitment to Deepen Work on Severe AI Risks Concludes AI Seoul Summit. <https://www.gov.uk/government/news/new-commitment-to-deepen-work-on-severe-ai-risks-concludes-ai-seoul-summit> Accessed: 2025-05-04.
  - [168] UNICEF. 2021. Policy Guidance on AI for Children. <https://www.unicef.org/innocenti/media/1341/file/UNICEF-Global-Insight-policy-guidance-AI-children-2.0-2021.pdf>
  - [169] United Nations. n.d. History of the Biological Weapons Convention. <https://disarmament.unoda.org/biological-weapons/about/history/> Accessed: 2025-05-04.
  - [170] U.S. Congress. 2024. S.4569 - TAKE IT DOWN Act, 118th Congress (2023-2024). Congress.gov. <https://www.congress.gov/bills/118th-congress/senate-bill/4569> Sponsored by Sen. Ted Cruz [R-TX], referred to the Senate Committee on Commerce, Science, and Transportation. Passed Senate..
  - [171] U.S. Copyright Office. 2024. Copyright and Artificial Intelligence. <https://www.copyright.gov/ai/>
  - [172] U.S. House Committee on Foreign Affairs. 2024. H.R. 8315 — Enhancing National Frameworks for Overseas Restriction of Critical Exports Act (ENFORCE Act). <https://foreignaffairs.house.gov/wp-content/uploads/2024/05/HR-8315.pdf> Accessed: 2025-05-07.
  - [173] U.S. House of Representatives. 2023. H.R.5628 - Algorithmic Accountability Act of 2023. <https://www.congress.gov/bills/118th-congress/house-bill/5628> Accessed: 2025-05-01.
  - [174] U.S. Nuclear Regulatory Commission. 2011. Final Safety Culture Policy Statement. Federal Register, 34773–34778 pages. <https://www.federalregister.gov/documents/2011/06/14/2011-14656/final-safety-culture-policy-statement> Document No. 2011-14656, Docket NRC-2010-0282.
  - [175] Battle v. Microsoft Corporation. 2023. <https://www.courtlistener.com/docket/67597344/battle-v-microsoft-corporation/> U.S. District Court for the Western District of Washington.
  - [176] John Villasenor. 2025. DeepSeek shows the limits of US export controls on AI chips. <https://www.brookings.edu/articles/deepseek-shows-the-limits-of-us-export-controls-on-ai-chips/> Accessed: 2025-05-07.
  - [177] Eugene Volokh. 2023. Large Label Models? Liability for AI Output. 489 pages. <https://ssrn.com/abstract=4546063>
  - [178] Hanna Wallach, Meera Desai, Nicholas Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2024. Evaluating Generative AI Systems is a Social Science Measurement Challenge. arXiv:2411.10939 [cs.CY] <https://arxiv.org/abs/2411.10939>
  - [179] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. arXiv:1905.00537 [cs.CL] <https://arxiv.org/abs/1905.00537>
  - [180] White House Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. <https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/> Accessed: 2025-05-23.
  - [181] White House Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights: Safe and Effective Systems. <https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/safe-and-effective-systems/> Accessed: 2025-05-01.
  - [182] Danae G. Widdler and Dawn Nafus. 2022. Dislocated Accountabilities in the AI Supply Chain: Modularity and Developers’ Notions of Responsibility. <https://arxiv.org/pdf/2209.09780v2.pdf>. Retrieved October 15, 2023.
  - [183] World Nuclear Association. 2021. Safeguards to Prevent Nuclear Proliferation. <https://world-nuclear.org/information-library/safety-and-security/non-proliferation/safeguards-to-prevent-nuclear-proliferation.aspx> Updated article from the World Nuclear Association Information Library, accessed 2025-05-09.
  - [184] Rui-Jie Yew. 2024. Break It ‘Til You Make It: An Exploration of the Ramifications of Copyright Liability Under a Pre-training Paradigm of AI Development. 9 pages. <https://doi.org/10.1145/3614407.3643707>
  - [185] Rui-Jie Yew, Bill Marino, and Suresh Venkatasubramanian. 2025. Red Teaming AI Policy: A Taxonomy of Avoidance and the EU AI Act. arXiv:2506.01931 [cs.CY] <https://arxiv.org/abs/2506.01931>
  - [186] Angela Zhang. 2025. US export controls have forced Chinese tech companies to be more innovative. <https://www.ft.com/content/c99d86f0-2d17-49d0-8dc6-9662ed34c831> Accessed: 2025-05-07.