Research Report

SASHA ROMANOSKY, ELINA TREYGER, ELIE ALHAJJAR

Legal and Policy Approaches to Mitigate Catastrophic Harms from AI

For more information on this publication, visit **www.rand.org/t/RRA4266-1**.

Published by the RAND Corporation, Santa Monica, Calif.

RR-A4266-1

About This Report

The dramatic rise of artificial intelligence (AI) and large language models (LLMs) has raised concerns about the potential for causing catastrophic harms. As AI models advance in capabilities, experts have warned of catastrophic risks to the United States caused by cyberattacks, bioweapons through direct malicious direction, and loss of control of these technologies. However, these risks remain uncertain and unquantified.

There has been no shortage of efforts—by governments, nongovernmental organizations, and AI developers—to create mandatory and voluntary guardrails, incentives, or other measures that would support AI safety and mitigate AI harms while still stimulating innovation. However, it is unclear which of these approaches would hold the greatest promise to prevent or reduce the probability of nationally catastrophic harms from AI systems.

Therefore, using a Delphi research methodology with 16 experts in AI and policy across the United States, we sought to identify what might be considered the most promising legal and policy measures that might reduce the probability of AI-caused catastrophic harms.

Center on AI, Security, and Technology

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Center on AI, Security, and Technology, which aims to examine the opportunities and risks of rapid technological change, focusing on AI, security, and biotechnology. For more information, contact cast@rand.org.

Funding

This research was independently initiated and conducted within the Center on AI, Security, and Technology using income from operations and gifts and grants from philanthropic supporters. A complete list of donors and funders is available at www.rand.org/CAST. RAND clients, donors, and grantors have no influence over research findings or recommendations.

Acknowledgments

We would like to acknowledge our colleagues who reviewed this document and provided useful comments. In addition, we would like to thank the participants of the online elicitation for devoting their time and effort to participate in our research study.

Summary

Sixteen experts participated in the online elicitation, which ran from January 9 to February 24, 2025. Experts consisted of artificial intelligence (AI) technology and policy researchers in academia or think tanks and senior government officials. The expert elicitation conducted for this study focused on assessing different categories of legal and policy measures that may reduce the probability of an AI-caused catastrophic harm. It ran for three rounds and produced the following findings:

- In general, participants from the online Delphi elicitation did not convey strong optimism about the potential for the 11 legal and policy categories presented to incentivize behavior that would reduce the probability of a catastrophic AI-caused harm: Most categories were assessed to be of uncertain desirability and feasibility.
- Participants became more skeptical about the feasibility of most of these categories by the end of the elicitation.
- In the views of the participants that emerged from the elicitation, the most promising legal and policy categories to shape incentives for AI developers to reduce the probability of catastrophic AI-caused harms were *incentives to find and disclose risks* and *voluntary safety standards*.
- Heavier-handed and stricter regulations, such as *mandatory audits*, *government-imposed restrictions on actors*, and *mandatory safety standards*, emerged as the least promising categories.

The perspectives from the expert elicitation imply the following:

- The likelihood of comprehensive measures being adopted by the federal government to reduce catastrophic AI risks in the near term (five to ten years) is perceived as low.
- Decisionmakers may be well-served by focusing on the more promising categories (such as those listed above), which may be implemented without federal government involvement.
- It may be more feasible for state governments to enact, and for industry and nongovernmental actors to adopt, some of the more promising measures.
- Although even the most promising categories have shortcomings, some shortcomings may be remediable in the near term, such as through appropriately structured disclosure programs and legal safe harbors for researchers.

Contents

Figures and Tables

Figures

Tables

Legal and Policy Approaches to Mitigate Catastrophic Harms from AI

The dramatic rise of artificial intelligence (AI) and large language models (LLMs) has raised concerns about the potential for causing catastrophic harms. As AI models advance in capabilities, experts have warned of "catastrophic risks unlike any the United States has ever faced" (Harris, Harris, and Beall, 2023, p. 5). The risks that advanced AI models pose are described as comparable to those of weapons of mass destruction; such models are described as capable of "executing catastrophic malware attacks, assisting in bioweapon design, and directing swarms of goal-directed humanlike autonomous agents" and executing "an untraceable cyberattack to crash the North American electrical grid" (Harris, Harris, and Beall, 2023, pp. 5, 26). Neither existential risks that threaten the survival of the species nor catastrophic harms have occurred, and the potential risks remain uncertain and contested. In light of these uncertainties, there has been no shortage of efforts by governments, nongovernmental organizations, and the largest AI developers to propose guardrails or measures that would support AI safety and mitigate AI harms. In the United States, most legislative efforts have not focused on catastrophic harms, often targeting the kinds of harms that have already materialized, such as discrimination, bias, privacy, and unethical uses of AI systems. Voluntary commitments by leading AI companies, voluntary standards for safe and secure AI proposed by the National Institute of Standards and Technology (NIST) and nongovernmental bodies, and numerous regulatory proposals cover a wider variety of AI-related harms, ranging from everyday harms, such as bias, to existential risks (see Cohen et al., 2024; Guha et al., 2024; and Stanford Institute for Human-Centered Artificial Intelligence [HAI], 2025, Ch. 6).

Which of the many measures featuring in legislative, regulatory, and policy debates might hold the greatest promise to prevent or reduce the probability of nationally catastrophic harms from AI systems? Because these risks are of harms that have not yet come to pass, there is little evidence to inform policymaking and little consensus on what measures should be adopted and by whom (e.g., Bommasani et al., 2024). Effective policymaking requires advances in the scientific understanding of the risks and the effects of different measures, which are likely to take time (Bommasani et al., 2024). At the same time, experts stress the urgency of acting now (Anderljung et al., 2023; Bengio, 2024).

Introduction

In our study, we took one approach to identifying the most-promising legal and policy measures that might reduce the probability of AI-caused catastrophic harms in the absence of systematic scientific evidence. We relied on an expert elicitation based on the Delphi methodology (a structured approach to gathering expert opinions through multiple rounds of questions and interaction) to collect data related to the following research questions:

- What categories of legal and policy measures are viewed as most *effective* at changing incentives for developers and deployers of AI and malicious and nonmalicious users?
- What categories of legal and policy measures are viewed as most *desirable* and most *feasible* at reducing the probability of catastrophic harm from AI, and why?
- Combining perceived effectiveness, feasibility, and desirability, what are the more-promising categories of legal and policy measures that might reduce the probability of nationally catastrophic AI harm?

- What might make the more-promising categories of legal and policy measures more effective, desirable, and feasible?

Ultimately, as prominent experts in the debate observe, effective reductions in the probability of catastrophic harms will require international action and coordination (Bommasani et al., 2024). Such reductions will also require sustained investment in research and scientific advancement. This study, however, was limited to identifying measures that may be implemented by U.S. actors in the near term (five to ten years).

Defining Key Concepts

For the purpose of this research, *AI* refers to foundation models, following Bommasani et al. (2022), who define it as "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" (p. 3). Foundation models are rooted in deep neural networks and supervised learning; once pretrained, foundation models can be fine-tuned or adapted for specific applications. A special class of foundation models is *frontier AI models*, which Anderljung et al. (2023) define as "highly capable foundation models that could exhibit sufficiently dangerous capabilities" (p. 7). While there is no comprehensive list of such dangerous capabilities, they encompass those that can produce mass casualties, physical harm, and societal disruptions, among others.

Such dangerous capabilities may lead to catastrophic harms. There is no singular, widely accepted definition of such harms, and definitions are context dependent.[1] We focus on harms that are catastrophic at the national level, and we avoid any strict quantification of the harm: We define *catastrophic harms* as **those that produce substantial loss of life or cause tremendous property damage of national significance** (see, e.g., Viscusi and Zeckhauser, 2012). Defining catastrophic harm in terms of concrete losses of life or money means that we do not consider more-dispersed and gradually unfolding harms, such as the erosion of democratic norms and institutions or the economic dislocations produced by the increasing role of AI systems in the national economy.[2]

AI systems might lead to catastrophic harms in several ways, which may be distinguished by reference to the intent of the human users of AI:

- **Accidentally caused harms:** These harms could result from misalignment between the goals pursued by the AI system and the goals of its human developers or users. Misalignment might happen through poor specification, reward hacking, or emergent goals that do not align with human goals or interests, leading to unintended outcomes, including catastrophic harms (e.g., Bommasani et al., 2022; Pan, Bhatia, and Steinhardt, 2022). These scenarios are particularly concerning in high-stakes and mission-critical environments (Cohen et al., 2024). At the extreme, and with highly capable systems, misalignment might lead to a loss of control, or scenarios in which AI systems acquire autonomy to pursue their own goals and humans cannot easily reassert control or shut down the systems (Cohen et al., 2024).
- **Maliciously caused harms:** Catastrophic harm may also arise when AI systems are exploited or weaponized by human actors with malign intentions. An AI system might be used, for example, to synthe-

[1] For example, the NIST Federal Information Processing Standard Publication 199 for protecting information systems defines a *catastrophic adverse effect event at an organizational level* as one that might "(i) cause a severe degradation in or loss of mission capability to an extent and duration that the organization is not able to perform one or more of its primary functions; (ii) result in major damage to organizational assets; (iii) result in major financial loss; or (iv) result in severe or catastrophic harm to individuals involving loss of life or serious life threatening injuries" (NIST, 2004, p. 3).

[2] That is not to say that such losses are not catastrophic—indeed, these types of harms may ultimately have greater national significance than even large-scale losses of life and property.

size existing knowledge to identify new opportunities for inflicting catastrophic damage, such as being employed to automate the discovery of software vulnerabilities in critical U.S. infrastructure, engineer bioweapons, orchestrate deceptive information campaigns, or deploy autonomous robots for harmful purposes (Brundage et al., 2024).

Even if current generations of AI systems may not possess the capabilities necessary to inflict catastrophic harm, some experts argue that, with further progress, the potential for causing significant harm could increase dramatically (Bengio et al., 2024; Uuk et al., 2024). These risks are compounded by the uncertainties about the capabilities of AI systems at any given point: Advanced AI models may possess significant latent capabilities that can remain undetected for years following their development and public release. The escalation in capability combined with uncertainty raises critical questions about the need for proactive measures to ensure the safe development and deployment of these technologies (Bostrom, 2014; Carlsmith, 2024). We do not take a position about the precise probability that AI systems can, in fact, cause catastrophic harms; however, we start from a presumption that the concerns noted above are sufficiently serious to warrant focused effort to prevent such harms.

The Policy and Legal Landscape

Efforts to mitigate the potential for AI-caused harms have taken various forms. In recent years, governments have increasingly recognized the need to regulate AI to mitigate its risks while leveraging its benefits. Internationally, the European Union (EU) has been the most active source of AI governance and regulation. Following on its 2018 General Data Protection Regulation (GDPR), it introduced the EU AI Act, the most comprehensive AI legislation globally (European Parliament, 2025). At least 39 other nations have passed some AI-related legislation (HAI, 2025, p. 17). These laws vary in scope and address a broad range of uses, concerns, and risks stemming from rapidly developing AI systems. Although risks of catastrophic harms are rarely identified as such, the EU AI Act, for example, does identify some AI systems as high risk.

In contrast to the EU, in the United States, federal action to mitigate risks of increasingly capable AI systems has been limited. Several narrowly focused laws have been passed, and a growing number of federal regulations that implicate AI in specific contexts have been issued (HAI, 2025). The White House's AI Action Plan sets out some recommended policy actions (such as export controls and AI evaluations) to guard against emerging risks while removing regulatory barriers to innovation, although it remains to be seen how these recommendations will be implemented (White House, 2025).

State-level activity has intensified: 49 laws were passed in 2023, and 131 more were passed in 2024 (HAI, 2025). State-level laws have also tended to focus on specific applications and risks, such as AI in law enforcement, privacy measures, consumer protection, and deepfakes (HAI, 2025). None focus on the risks of catastrophic harms or a broader set of risks that includes such harms. The one state bill that did explicitly address catastrophic harm, California Senate Bill 1047 (2024), drew controversy and was vetoed by the governor (Shepardson and Tong, 2024). Colorado's legislation on high-risk AI systems may be relevant to mitigating catastrophic harms but does not define *high-risk* in terms of such harms. New York's Responsible AI Safety and Education Act, or RAISE Act, seeks to drive AI developers to implement appropriate safeguards to prevent "critical harms" and holds developers civilly liable for such harms.[3]

Along with private-sector measures, efforts by governmental entities, such as the United Kingdom (UK) AI Security Institute and NIST (which houses the Center for AI Standards and Innovation [CAISI]), have

[3] In this act, *critical harm* is defined as "the death or serious injury of one hundred or more people or at least one billion dollars of damages to rights in money or property caused or materially enabled by a large developer's use, storage, or release of a frontier model" (New York State Assembly Bill 6453, 2025, para. 7).

produced a host of voluntary measures aimed at safe and secure development of AI. Notably, as a result of the UK AI Safety Summit in 2023 and the AI Seoul Summit in 2024, several nations and major AI developers agreed to a set of principles and cooperation aimed at managing the risks of AI and its responsible development and deployment. NIST has developed an AI Risk Management Framework, which addresses a broad set of risks, including those of severe or catastrophic harms (NIST, 2023).

Literature on Mitigating AI-Caused Catastrophic Harms

Academic and policy research and discussions of AI-related risks and the efforts to mitigate them are growing rapidly. Efforts to focus specifically on the potential for catastrophic harms and how these might be prevented or mitigated are expanding: The Future of Humanity Institute (undated), Future of Life Institute (undated), Center for AI Safety (undated), and Centre for the Study of Existential Risk (undated) are among the organizations working on addressing catastrophic harms from AI. While some studies touch on the potential for AI to create such risks—for example, concerning the development of autonomous weapons or misaligned superintelligence (Bengio et al., 2024; Bostrom, 2014; Carlsmith, 2024)—comprehensive strategies for preventing these catastrophic outcomes remain underexplored. A growing body of research analyzes the effects of particular measures or sets of measures for mitigating harm, such as regulatory frameworks, auditing practices, and liability regimes (Beckers and Teubner, 2022; Guha et al., 2024; Weil, 2024).

Because empirical evidence on the effects of different measures on the risks of AI harms is unavailable, researchers have sought to synthesize expert appraisals of such measures. A few researchers have surveyed experts on their views on long menus of measures that might reduce the risks of significant harms or otherwise produce safer AI. Notably, Uuk et al. (2024) surveyed AI experts to assess the perceived effectiveness and technical feasibility of many mitigation measures for four sets of risks: disruptions to critical sectors; harm to democratic processes; chemical, biological, radiological, and nuclear (CBRN) risks; and harmful bias and discrimination. The researchers presented measures that would be relevant to implementing the EU AI Act, presuming that each measure could be implemented fully.

Similarly, Schuett et al. (2023) surveyed experts to assess which of the many safety and governance practices are desirable for AI labs to voluntarily adopt. Both Uuk et al. (2024) and Schuett et al. (2023) found expert consensus that a large number of the measures that were considered are effective and technically feasible and desirable or should be adopted by AI labs: Uuk et al. (2024) found that a broad selection of the 27 mitigation measures presented were perceived as effective and technically feasible by experts.[4] Schuett et al. (2023) found strong expert agreement that all but one of the 50 presented actions should be implemented by AI labs.

These are informative results. But they are difficult to translate into actionable priorities, especially for U.S. decisionmakers and policy experts concerned about reducing catastrophic risks. First, a long wish list of measures—especially those that require government action and extensive government effort and resources—is less practically feasible in the U.S. context than in the EU in the near future. Second, effectiveness and technical feasibility need to be weighed against the potential costs and unintended adverse effects of each measure. A measure that is technically feasible and highly effective at reducing risk may still bring significant unintended effects—and should not be prioritized for that reason. Third, measures that emerge as best practices for voluntary adoption may be less desirable and/or not feasible as government mandates. Identifying

[4] More specifically, Uuk et al. (2024) also found that three mitigation measures stood out for having the highest expert agreement ratings across all risk areas and being the most frequently selected in experts' preferred combinations of measures: safety incident reports and security information-sharing, third-party predeployment model audits, and predeployment risk assessments.

which of the available measures are most promising to reduce the probability of the more severe harms that AI can bring about is therefore necessary to enable proponents of AI safety and security to focus their efforts on the most effective strategies for risk mitigation.

Scope and Organization of This Report

In this report, we seek to fill some of the gaps in the existing research. That is, we seek to identify a tractable set of legal and policy measures that might be effective at reducing the probability of nationally catastrophic harms (relative to other available measures), have relatively few adverse effects, and are feasible to implement in the United States in the near term. To do so, we focused on a subset of all possible legal and policy measures: those that could reduce the probability of catastrophic harms *through incentivizing one or more of the key actors in the causal chain leading to the harm to behave in ways that may reduce the likelihood of catastrophic harm*.[5]

Furthermore, we consider that catastrophic harms may result through one of three notional mechanisms: accident, malicious action, and autonomous AI or loss of control. For all three mechanisms, **AI developers** are one type of key actor whose actions can affect the probability of catastrophic harm. AI developers are most well-informed about the capabilities and vulnerabilities of their models—and are therefore best positioned to take precautionary measures that reduce the probability of harms. The opportunity for shaping the probability of harm is what places AI developers in the causal chain from action to a hypothetical harm. For harms caused through malicious actions, **malicious users** are another key actor: By definition, a malicious action requires a malicious actor to act. And for harms caused accidentally—that is, without ill intent—**nonmalicious (i.e., benign) users** may be another key actor. Accidental harms may also occur without any users at all, as in a loss-of-control scenario, in which case the AI developer may be the sole key actor. Reducing the set of key actors in these causal chains to harm is doubtless an oversimplification, but one that is needed to produce a tractable comparison of legal and policy measures.[6]

Focusing on policies and laws that shape incentives meant that we did not consider such options as research and development (R&D) funding, which might create better safety and security mechanisms; increased investment into AI safety or security generally; preparedness and response measures to mitigate harms after they occur; and a host of other measures. Preparedness and response measures to mitigate harms after they occur may well reduce the likelihood that AI causes harm or reduce the magnitude of the harm if it occurs, but primarily not through materially shaping the incentives of key actors to develop or use AI systems in a manner that reduces that likelihood.[7]

The rest of this report is organized as follows: The next section describes our research methods. We then present a summary of findings, followed by key insights derived from our qualitative and quantitative elicitation exercise. We conclude with policy implications for AI governance in the United States.

[5] We consider risk to be a function of the probability of harm multiplied by the magnitude of harm, and so in this context, because we are focused on catastrophic harm only, we consider policies that focus on reducing the probability of harm. We recognize, of course, that some policies may reduce both probability and magnitude in varying amounts.

[6] We do not, for example, distinguish between malicious and nonmalicious developers or include victims.

[7] That is, R&D into safe AI can produce technological breakthroughs that enable a safer AI; however, the call to increase R&D itself does not incentivize any particular actors to undertake or refrain from any particular course of action.

Research Methods

Identifying the more promising types of policies or legal measures to reduce the probability of a harm that has never materialized (to date) is challenging. There is no empirical record to examine and few, if any, close historical precedents to draw on. There is great uncertainty about the precise mechanisms whereby frontier or future AI systems may produce catastrophic harms, the magnitude of those harms, the probability of them occurring, and, therefore, the effects of any potential measures to reduce that probability.

The Delphi method is particularly useful under these conditions. The *Delphi method*, as RAND researchers explain,

> is an iterative, anonymous, structured, group-based communication process and elicitation technique designed to help policymakers make decisions under conditions of uncertainty and incomplete information. It is based on the premise that asking a hand-picked group of anonymous experts the same questions several times and sharing the other experts' answers will help objectively develop group consensus, which is used as a form of evidence. (Khodyakov et al., 2023, back cover)[8]

The Delphi method is most instructive when used to illuminate structured expert consensus under conditions of strategic uncertainty and data scarcity, exactly the context of frontier AI risks. It helps distill areas of agreement and contention among domain experts regarding plausible threat models, intervention points, or governance strategies, particularly when these are contested or under-theorized. For example, it can identify which policy levers are perceived to have the broadest support or which risk scenarios experts consider most urgent, thus informing early agenda-setting or prioritization by decisionmakers. It is especially valuable when the goal is to surface judgment-based heuristics rather than empirical prediction.

We assembled a group of experts and asked them to assess different categories of legal and policy measures that may reduce the probability of an AI-caused catastrophic harm. We sought to assemble a set of participants with varied backgrounds and from various sectors to increase the likelihood that they would not all be drawing on the same set of assumptions or domain of knowledge.

Here, we first describe participant backgrounds, and then we explain (1) how we reduced the large number of legal and policy measures into a tractable number, (2) how we defined and measured the effectiveness, feasibility, and desirability of these categories, and (3) how the elicitation itself was designed and conducted.

Participant Backgrounds

We invited 42 individuals to participate, and 24 accepted. The elicitation consisted of three rounds of engagement and was conducted using RAND's ExpertLens web-based platform, which implements a modified Delphi methodology (Khodyakov et al., 2023). Participant selection was based on either their technical expertise in AI systems or their experience in technology policy or legal domains (including AI). Three of the 24 individuals were unable to complete round 1 and were removed from the elicitation. Of the 21 participants

[8] Khodyakov et al. (2023, back cover) also note,

> Although it was originally developed by RAND researchers as a forecasting methodology in military research, Delphi underwent many modifications and is now used by different disciplines, most notably by medicine, as a gold-standard approach for expert elicitation and stakeholder engagement. Researchers often rely on Delphi to estimate the probability of an event happening within a certain period of time, to forecast when an event is likely to occur, and to identify and prioritize key policy issues that need to be addressed.

who began the elicitation, 20 completed round 2, and 16 completed round 3. The elicitation started January 9 and ended February 24, 2025. The participants had the following backgrounds:

- Of the 24 participants who agreed to participate, 21 held senior positions in their respective organizations with at least ten years of industry or research experience; the three participants with less experience were experts specifically in AI technology and policy.
- The participants were policy and technology researchers in academia or think tanks (14), private-sector experts in technology and/or law and policy (six), and senior government officials (four).
- The academic participants were senior researchers in technology, policy, law, and/or AI.
- The industry participants came from cybersecurity, insurance, and law (with those from the latter two industries focused on technology and/or AI).
- Government participants came from federal agencies and held senior policy and technology roles.

Legal and Policy Measures to Reduce Probability of Catastrophic Harms

As noted earlier, this study focused on the subset of legal and policy measures that could reduce the probability of catastrophic harms *through incentivizing one or more of the key actors to behave in ways that reduce the likelihood of catastrophic harm*.

We derived our categories or sets of legal and policy measures from two sources. First, we considered proposals that have been made or adopted to reduce the likelihood of AI harms specifically. Although some, or even most, of these measures were not geared exclusively or primarily at catastrophic harms, they are viewed as reducing risks of all sorts of AI harms.[9] Second, we considered measures that have historically been used to incentivize actors to reduce the probability of other mass and catastrophic harms. Historically, preventing or reducing the likelihood of catastrophic harms has been pursued through a combination of ex ante and ex post legal and policy measures. *Ex ante* measures, such as safety regulations, seek to reduce the probability of the harm (or prevent the harm entirely). *Ex post* measures, such as tort law, criminal law, and insurance, seek to compensate for the losses and/or penalize parties who might be responsible for the harms—and, thereby, incentivize parties to take precautions to minimize the risks (reduce probability of harm) ex ante.

There are many ways to categorize the policies and legal measures. In this study, we sought to group measures that could be described by reference to (1) a common incentive-based logic, including whether the measure is one that applies ex ante or ex post, and (2) common policymakers or decisionmakers who would impose or create a measure within each set. Thus, some categories consisted of laws or regulation mandates by some government actors, some consisted of measures voluntarily adopted by private or nongovernmental actors, and some were amenable to implementation by a variety of decisionmakers. Categories that rely on government decisionmakers also implicate different branches of government. For example, measures labeled as *mandatory* tend to require action from legislators or regulatory agencies, while other categories, such as those involving tort or criminal liability, depend more heavily on the courts for implementation and enforcement.

In the pre-elicitation step, we engaged ten policy experts at RAND to help identify the appropriate number, framing, and grouping of 11 categories of policy and legal measures that we generated based on the above considerations. We revised the descriptions of the categories in response to feedback provided by these experts.[10] These 11 categories are presented in Table 1, and a complete description is provided in Appendix B.

[9] We excluded measures that narrowly targeted specific, noncatastrophic harms—such as bias or deepfakes—from consideration.

[10] These experts provided comments, and we held conversations with four of them to further elaborate on their topics. We then made minor revisions and improvements (no major or material changes) to a few of the policy descriptions. The complete prompt is presented in Appendix A.

TABLE 1

Legal and Policy Measures

Policy Category	Description	Policymakers or Decisionmakers
Ex ante measures		
Mandatory risk assessment and mitigation frameworks and processes	Legal requirements for developers to adopt internal governance structures or frameworks or institute internal processes to reduce catastrophic harms (e.g., chief risk officers, risk assessments)	• State or federal governments • Legislators or regulators
Mandatory safety and security standards	Standards, such as secure development, deployment, and/or use, to prevent unauthorized or malicious use (e.g., curation of the training data, curation of requests and responses, kill switch, protection of model weights)	• State or federal governments • Legislators or regulators
Voluntary safety and security standards	Voluntary implementation of safety and security standards described above	• Industry • Standard-setting or self-regulatory organizations
Mandatory monitoring and audits	Mechanisms for verifying that an AI model is performing as expected; compliance with legal requirements or safety standards (e.g., monitoring, audits, or evaluations of AI systems)	• State or federal governments • Legislators or regulators
Voluntary monitoring and audits	Voluntary implementation of the category described above	• Industry • Standard-setting or self-regulatory organizations
Restrictions on capabilities or applications	Prohibition of particular capabilities, applications, or uses of AI systems (e.g., weapon development, biological, cyber)	• State or federal governments • Legislators or regulators • Industry • Standard-setting or self-regulatory organizations
Restrictions on actors	Limitations or regulations on which actors are allowed to develop, deploy, and/or use AI systems (e.g., licensing, regulatory sandboxes, export controls, Know Your Customer guidelines and regulations)	• State or federal governments • Legislators or regulators
Incentives to find and disclose risks	Measures that create incentives to expose vulnerabilities or risky or unlawful practices (e.g., whistleblower protections, bug bounties or vulnerability disclosure programs, contests and challenges)	• State or federal governments • Legislators or regulators • Industry • Standard-setting or self-regulatory organizations
Mandatory disclosure	Requirements to share information about AI risks, vulnerabilities, or aspects of its performance and safety (e.g., AI registry, risk or impact assessments, security and safety incident reporting)	• State or federal governments • Legislators or regulators
Ex post measures		
Tort liability	Enablement of private individuals or government agents to bring legal actions against injurers (developers, deployers, or users)	• State or federal governments • Courts • Private parties
Criminal liability	Empowerment of government agents to bring criminal charges against injurers (developers, deployers, or users)	• State or federal governments • Courts

NOTE: Categories that consist of the same measures, distinguished by their mandatory or voluntary character, are presented one after another.

Measurement Dimensions

We asked our panel of experts to evaluate these categories—which we describe as *policy sets*, for brevity—along three dimensions that are relevant to assessing their overall promise for successful implementation (see Table 2). First, we asked experts to assess the effectiveness of each set, or the extent to which it incentivizes each of the key actors (developers, nonmalicious users, and malicious users) to act in ways that reduce the probability of a nationally catastrophic harm. Second, we asked experts to assess the **feasibility** of implementing the measures from a given category—in the United States and in the near future (approximately five to ten years). *Feasibility*, for the purposes of this study, was defined in terms of any technical, practical, legal, political, or other factors that affect the chances of its implementation in practice (as was explained to the participants). Third, we asked experts to assess the overall desirability of implementing the policy. *Desirability* in this context is a net assessment of both advantages and drawbacks or costs of policies in a given category, which may include considerations related to innovation, competition, privacy, security, cost, environmental, or other factors. For each dimension, we asked the expert participants in our elicitation to make an assessment on a Likert scale of 1 to 5, in which each numerical score corresponded to a (slightly different) qualitative assessment. We summarize these dimensions and scoring scales in Table 2.

Elicitation

In round 1, participants were presented with the list of 11 policy categories to consider and assign a score to each dimension, following the definitions set out in Table 2. For feasibility and desirability assessments, participants also provided free-form comments to explain their assessments. Participants were not asked for comments on their effectiveness assessments; because effectiveness with regard to each key actor was assessed separately, commenting on each would have substantially lengthened the time required to complete the round. Although we did not solicit comments on effectiveness assessments directly, because effectiveness is a likely factor in assessing overall desirability, participants had the option to—and many did—comment on effectiveness in that part of the elicitation.

TABLE 2
Three Assessment Dimensions and Scales

Dimension	Assessment Question and Scale
Effectiveness	How effective would this set of policies be at incentivizing each of the following actors in a way that reduces the probability of a nationally catastrophic harm? [Actors presented were developers, nonmalicious users, and malicious users. A separate effectiveness score was solicited for each of the key actors.]
	1 = Counterproductive, 2 = Likely not effective, 3 = Slightly effective, 4 = Moderately effective, 5 = Highly effective
Feasibility	How feasible do you think it would be to implement this set of policies in the United States in the near future (e.g., five to ten years)? In your response, you may consider technical, practical, legal, political, or other factors.
	1 = Not feasible at all, 2 = Not very feasible, 3 = Uncertain, 4 = Moderately feasible, 5 = Highly feasible
Desirability	How desirable do you think it would be to implement this set of policies? In your response, you may want to consider pros and cons of this set of policies for innovation, competition, privacy, security, cost, environmental, or other factors.
	1 = Not desirable at all, 2 = Not very desirable, 3 = Uncertain, 4 = Moderately desirable, 5 = Highly desirable

Once participants had completed round 1, we summarized participants' free-form comments by policy category using RAND's private instance of Microsoft Azure–based OpenAI,[11] followed by our own review for accuracy and completeness.[12]

In round 2, the 20 participants who participated were presented with the distributions of Likert scores for effectiveness (three distributions, one for each actor), feasibility, and desirability for each set—that is, they were presented with the percentage of respondents who picked each score for each question and where their own answer fit in that distribution.[13] They were also presented with summaries of free-form comments and the complete text of all comments (to read if they chose to). Participants were asked to review others' scores and summary comments and to consider their answers relative to others. They were invited to explain what factors were important in their assessments and to respond to points raised by others that were different from their own thinking. Participants were invited to engage in an anonymous, asynchronous, moderated discussion board. Anonymity is important to the design because it allows participants to consider others' arguments and comments on their merits, unaffected by the identity of the authors.

In the third round, participants were presented with the same question(s) as in round 1 and were asked to consider updating their scores based on discussions from round 2. Although the purpose of this approach is typically to reach *consensus* (across all participants) concerning a given topic, we did not seek to reach consensus. However, we did encourage participants to update their scores based on considerations raised by others, which they may not have considered previously. We relied on one method (among several possible) to synthesize and summarize scores and comments so as to capture both initial and updated views: In our summary of findings, we further detail the choices we made to analyze and summarize the scores.

Limitations

Our methodological approach has limitations. First, our results are based on insights from a small sample of participants. Participants are knowledgeable about AI and policy and have varied backgrounds and experiences, which lowers the chances of systematic biases across participants. However, insights may still be biased toward the experiences of this sample group and potentially not generalizable beyond this group. For example, this sample of respondents may have been especially pessimistic about federal legislation, whereas another group would not have been (although we have no reason to believe that this was the case). Moreover, while convergence of expert opinion can signal robustness, it may also suppress minority or contrarian views, which are often vital in emerging technologies with unknown unknowns.

Second, not all experts participated in all three rounds of the ExpertLens elicitation. Although we believe that this did not significantly affect results, it is a limitation of the elicitation.

Third, to offer a practically manageable number of policy sets, each category encompassed a potentially wide variety of legal and policy measures. Thus, our experts' overall quantitative assessments obscured any significant differences that they perceived among different policies within a category.

Fourth, although we sought to generate a ranking of policy categories from most to least promising in the present context, the differences between some of the categories were relatively minor. Moreover, alternative methods for ranking might produce slightly different rankings, as we discuss later in the report. Because there is no singular *best* way to rank options across multiple dimensions, the specific results we highlight are

[11] The summary content was generated using model gpt-4o-2024-05-13 and a temperature of 0.0.

[12] We verified that the summaries matched our qualitative review of the comments and made adjustments as needed for greater accuracy.

[13] One of the 21 participants in round 1 did not participate in round 2.

a function of choices we made in this respect. In addition, this method is not designed to identify the marginal effectiveness of each policy category over another.

Finally, it is reasonable to believe that AI capabilities, as well as the landscape of politics and policies, are changing rapidly, and so while the threat of catastrophic harms may persist, techniques and political appetite for managing these risks will evolve over the near term. Experts may well come to different conclusions in future elicitations.

Summary of Findings

Table 3 presents a summary of all the quantitative assessments—the average Likert scores across each scoring dimension (effectiveness for each of the actor groups, feasibility, and desirability) for each policy category—that were produced in the final round, after participants updated their scores.[14] Because not all participants completed all three rounds, these results reflect the assessments of 16 of the original 24 participants.

TABLE 3
Average Likert Scores, Round 3

Policy	Desirability	Feasibility	Effectiveness for Developers	Effectiveness for Nonmalicious Users	Effectiveness for Malicious Users
Mandatory risk assessment	4.1	2.8	3.3	3.2	1.7
Mandatory safety standards	3.5	2.5	3.7	3.2	2.5
Voluntary safety standards	3.5	4.1	2.8	2.8	2.1
Mandatory audits	3.5	2.4	3.3	2.5	2.0
Voluntary audits	3.3	4.2	2.8	2.5	2.0
Restrictions on capabilities	3.4	2.9	3.3	3.4	2.7
Restrictions on actors	3.2	2.6	2.9	2.9	2.1
Incentives to disclose risks	4.0	3.9	3.6	3.1	2.5
Mandatory disclosure	3.9	3.2	3.1	2.8	1.8
Tort liability	4.0	2.7	3.6	3.5	2.9
Criminal liability	3.5	2.6	3.5	3.3	3.0

NOTE: Values represent the average of all Likert scores for all participants (*N* = 16) in round 3. Colors indicate relative scoring, with green colors representing *higher* scores on a scale of 1 to 5 across all values in the table and red colors representing *lower* scores. Feasibility is scored as follows: 1 = Not feasible at all, 2 = Not very feasible, 3 = Uncertain, 4 = Moderately feasible, 5 = Highly feasible. Desirability is scored as follows: 1 = Not desirable at all, 2 = Not very desirable, 3 = Uncertain, 4 = Moderately desirable, 5 = Highly desirable. Effectiveness is scored as follows: 1 = Counterproductive, 2 = Likely not effective, 3 = Slightly effective, 4 = Moderately effective, 5 = Highly effective. Values greater than 3 are presented in shades of green, values less than 3 are presented in shades of red, and values equal to 3 are neutral colored.

[14] In the course of this analysis, we examine both mean and median values. We present here the mean Likert values for the purpose of this discussion, and we show the median values in Appendix C. We use mean values because we assessed them to be the most appropriate way to summarize the group's view based on a five-point Likert scale, and this approach has been shown to be appropriate for such analysis (Norman, 2010). Analysis using median values greatly reduces the ability to reveal variation in scores (which, in our sample, is limited to 2, 3, and 4). Moreover, median values are most appropriate for highly skewed distributions, which did not appear in our data.

In terms of desirability, all of the categories considered were assessed to be somewhat better than *uncertain* but short of *highly desirable*, with none emerging as *not desirable* (i.e., none were below a 3 on average). By contrast, many of the categories were deemed *not very feasible* or of *uncertain feasibility* on average (i.e., scoring below a 3). Average effectiveness assessments varied across the three sets of key actors. Most categories were more effective for developers than users; for deployers, most averaged at or above *slightly effective* (i.e., scoring at least a 3), whereas for users, most were in the *likely not effective* range.

We dive deeper into the more salient conclusions from these scores below; but first, we present a summary of the comments made by participants regarding the feasibility and desirability of each of the policy categories, shown in Table 4.

A few general themes stood out in comments across categories and dimensions. Participants frequently noted political infeasibility, especially of mandatory measures. Unsurprisingly, voluntary measures were deemed more feasible in general, but many participants noted concerns with *safety-washing*—that is, prioritizing the appearance of compliance with measures over actually improving safety and reducing risks. Industry resistance to some measures (because of compliance burdens) and concerns about slowing or preventing innovation were also prominent, more so for the mandatory measures. When it comes to applying background principles of law (i.e., tort and criminal liability), participants expressed some skepticism that large, corporate actors would be deterred.

While we do not aim to elaborate on every theme that emerged from the elicitation, in the following section, we focus on the key insights that emerged from both the quantitative and qualitative results.

TABLE 4
Summary of Participant Insights

Policy Category	Summary
Mandatory risk assessment and mitigation frameworks and processes	• Political divisions significantly affect discussions on AI risk management, creating barriers between regulation and innovation. • Effective AI risk management frameworks face challenges because of the absence of established standards and reliable risk assessment methods. • Although this policy set is technically feasible, successful implementation relies on stakeholder support and may encounter issues, such as high costs and superficial compliance.
Mandatory safety and security standards	• Overall, there is a lack of political consensus to force implementation of safety standards, as well as resistance from the technology companies. • This category has reduced effectiveness to deter criminal actors, given the prevalence of open-source models. • There is broad agreement on the desirability for such a policy but disagreement about what would constitute proper safety or security controls.
Voluntary safety and security standards	• Lack of enforceability and risk of safety-washing are potential limitations. • This category could present an opportunity to foster consensus among industry regarding risk-mitigating controls. • Potential exists for voluntary standards to preempt stricter mandated standards.
Mandatory monitoring and audits	• The logistics of mandatory auditing pose difficulties because of the lack of structured frameworks, capacity, and technical expertise, requiring years of development for viability. • Significant political opposition and legal challenges, including concerns over trade secrets and industry lobbying, complicate the imposition of mandatory audits. • Although audits are crucial for ensuring system integrity and trust, their high costs and potential to stifle innovation raise concerns, particularly for smaller companies.
Voluntary monitoring and audits	• Voluntary measures are favored for their ease of adoption and appeal to tech developers compared with mandatory regulations. • Voluntary initiatives may lack effectiveness because of opt-outs, self-reporting issues, and potential manipulation by companies to enhance their reputations. • There is a consensus that establishing mandatory standards is essential for meaningful compliance, although voluntary audits can still provide initial risk-informing benefits.

Table 4—Continued

Policy Category	Summary
Restrictions on capabilities or applications	• Coordination across countries will present difficulties because of varying political priorities. • Restrictions may apply only to niche areas or actors. • These restrictions are a sound approach but difficult to enforce, given the uncertainty of AI systems.
Restrictions on actors	• The rapid pace of development makes these kinds of restrictions difficult. • Strict regulations could drive legitimate innovators to other countries (or underground), beyond the reach of any regulation. • Licensing the use could be an effective way to monitor the activity of high-risk AI systems.
Incentives to find and disclose risks	• These are low-cost ways of producing information about risks that already exist in other domains, but some participants were skeptical about actual reduction in catastrophic (versus daily) risks. • These measures may allow malicious actors to exploit vulnerabilities. • Some developers may resist, and government interest for whistleblower protection may be low.
Mandatory disclosure	• Companies may resist disclosure requirements because of concerns over costs, competitive sensitivities, and a lack of expertise or willingness to comply. • The feasibility of disclosure requirements is affected by the political climate, and challenges include ensuring the collection of useful information and consistent enforcement across jurisdictions. • While disclosure policies can enhance risk management and accountability, they may also lead to potential exploitation of vulnerabilities and could stifle innovation or impose compliance burdens.
Tort liability	• Default tort law already applies, but courts lack expertise to apply law to this new technology. • Tort liability can provide incentives for safer AI development, but companies will use contracts and insurance to reduce their tort liability risks. • While tort liability can encourage safer practices, it might also increase costs and legal burdens, particularly for smaller developers.
Criminal liability	• Criminal liability could deter malicious uses and worst abuses by developers, but it has not been clearly effective in deterring harms by corporations in the past; clear precedent would have to be set. • Opposition from developers to new laws is likely (but likely less than opposition to tort liability); it may be more feasible to apply to malicious users. • The potential for pushing out smaller companies and the likelihood that big companies will find loopholes exist.

NOTE: Summaries extract the most-common themes from participant comments on desirability and feasibility in rounds 1 (N = 24) and 3 (N = 16). Although participants did not comment on effectiveness specifically, comments on desirability often included comments on effectiveness as one factor that weighs in the overall desirability assessment.

Key Findings

In General, Responses Did Not Convey Great Optimism About the Feasibility or Effectiveness of the Legal and Policy Categories Presented

Overall, most categories were perceived as desirable to some degree but neither feasible nor very effective. As Table 3 shows, only two to three categories—consisting of voluntary measures or measures that could be voluntary (incentives to disclose risks)—were viewed as moderately feasible on average (scores close to or greater than 4). The rest were seen as uncertain in terms of feasibility at best. Desirability scores were somewhat higher, with four policy categories assessed as moderately desirable (score of 4) on average across participants, with the remainder closer to uncertain (scoring 3.5 or less).

None of the policy categories were viewed as moderately effective on average (a score of 4) for developers, with most categories scoring closer to slightly effective (scoring 3.5 or less), as shown in Table 3. Similarly,

effectiveness scores were highest for developers (the average score across all policy categories was 3.3, or just above slightly effective), followed by nonmalicious users (average score of 3.0), and lowest for malicious users (average score of 2.3).

The fact that most policy categories were deemed more effective for developers than other actors is not unexpected. The policy proposals to reduce the probability of AI-caused harms that feature prominently in public debates and research are overwhelmingly focused on the companies developing advanced AI models—which are reasonably perceived to have the greatest control over AI capabilities and postdeployment behaviors. To be sure, the policies we considered may also be applicable to users and shape their incentives: For example, safety and security standards, licenses, and tort and criminal liability may apply to user actions and developer actions. However, this is usually not the primary intended effect behind policy proposals, which likely accounts for our experts' assessments.[15]

In sum, participants did not have much confidence that any of these policy categories—taken individually—would be effective at incentivizing behaviors that reduce catastrophic risks or feasible to implement. This stands in some contrast to other studies that surveyed experts to assess some dimension of various legal and policy measures to improve AI safety or reduce certain risks. Although the research questions are not directly comparable across studies, these studies found that a large number of legal and policy measures (indeed, a vast majority of the 27 presented) were effective and technically feasible (Uuk et al., 2024), and 49 of 50 measures presented were deemed desirable and should be adopted by AI labs (Schuett et al., 2023).

Participants Became More Skeptical About Policy Feasibility by the End of the Elicitation

Many participants changed their scores from the first to the third round. This was expected and, indeed, part of the goal of the Delphi method. In a context rife with uncertainties and unknowns, arguments and considerations introduced by others may be expected to be more persuasive, which should lead to some revision. And many participants noted explicitly that they adjusted their scores based on thoughts offered by other participants. The direction of the changes between the first and last rounds, however, was overwhelmingly downward and primarily a result of the change in political administration and its appetite for AI regulation (discussed more below).

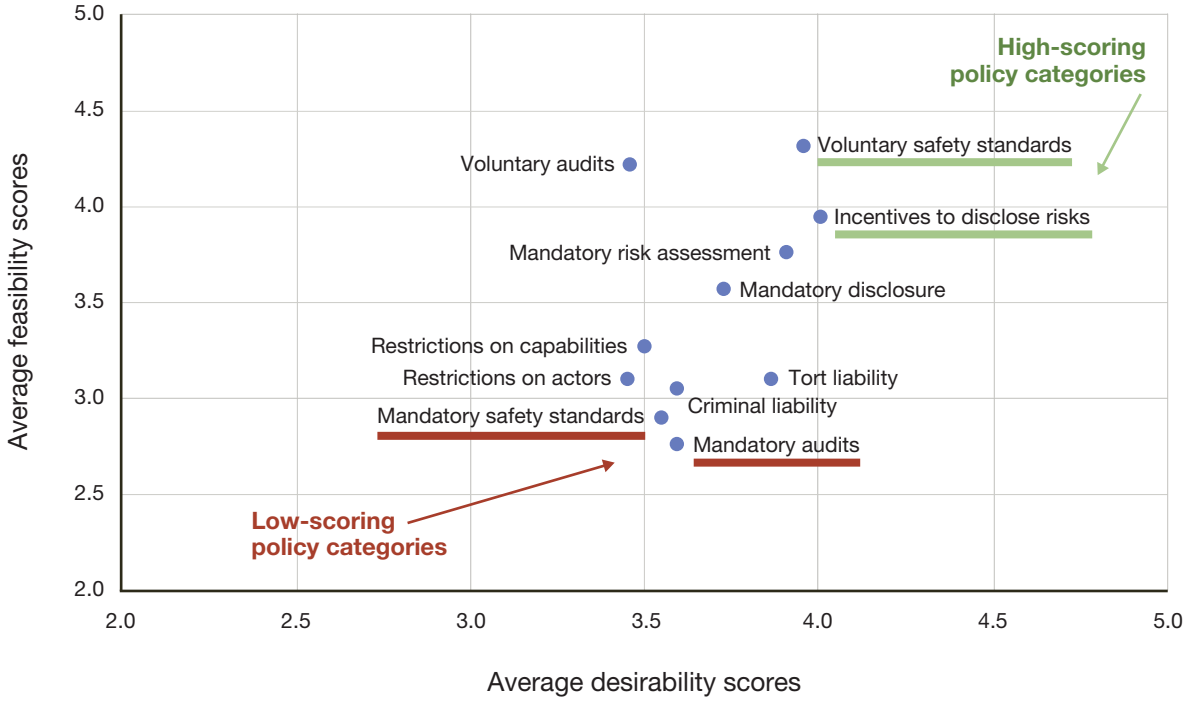
A scatterplot showing the desirability and feasibility scores for each policy category measured on a five-point Likert scale is shown in Figure 1.

In round 1, *voluntary safety and security standards* and *incentives to find and disclose risks* were considered most feasible and desirable. *Voluntary audits* were also assessed as among the most feasible, but less desirable than these other two categories: As one participant noted, "voluntary standards are highly feasible because they do not require the political will to impose penalties for nonadherence."

On the other hand, *mandatory audits* and *mandatory safety and security standards* were considered least feasible, though somewhat desirable. One participant commented that "mandatory audits would require standards against which companies would be audited but would be highly invasive and therefore undesirable to AI companies." Others emphasized technical uncertainty that makes it difficult to identify or comply with mandatory standards, saying, "Our knowledge regarding many safety and security challenges . . . is nascent," and "technically, there might also be significant uncertainty as to what it would mean to 'comply' with AI safety standards that are themselves uncertain and subject to change." Similar concerns were raised for man-

[15] In the pre-elicitation stage, we specifically asked experts to think about relevant legal or policy measures that are aimed at users who we might be omitting. No such measures were identified.

FIGURE 1
Desirability and Feasibility Scores in Round 1



NOTE: Scores represent the average desirability and feasibility scores across all participants for round 1 (N = 21), based on five-point Likert scales. The feasibility scale is as follows: 1 = Not feasible at all, 2 = Not very feasible, 3 = Uncertain, 4 = Moderately feasible, 5 = Highly feasible; and the desirability scale is as follows: 1 = Not desirable at all, 2 = Not very desirable, 3 = Uncertain, 4 = Moderately desirable, 5 = Highly desirable.

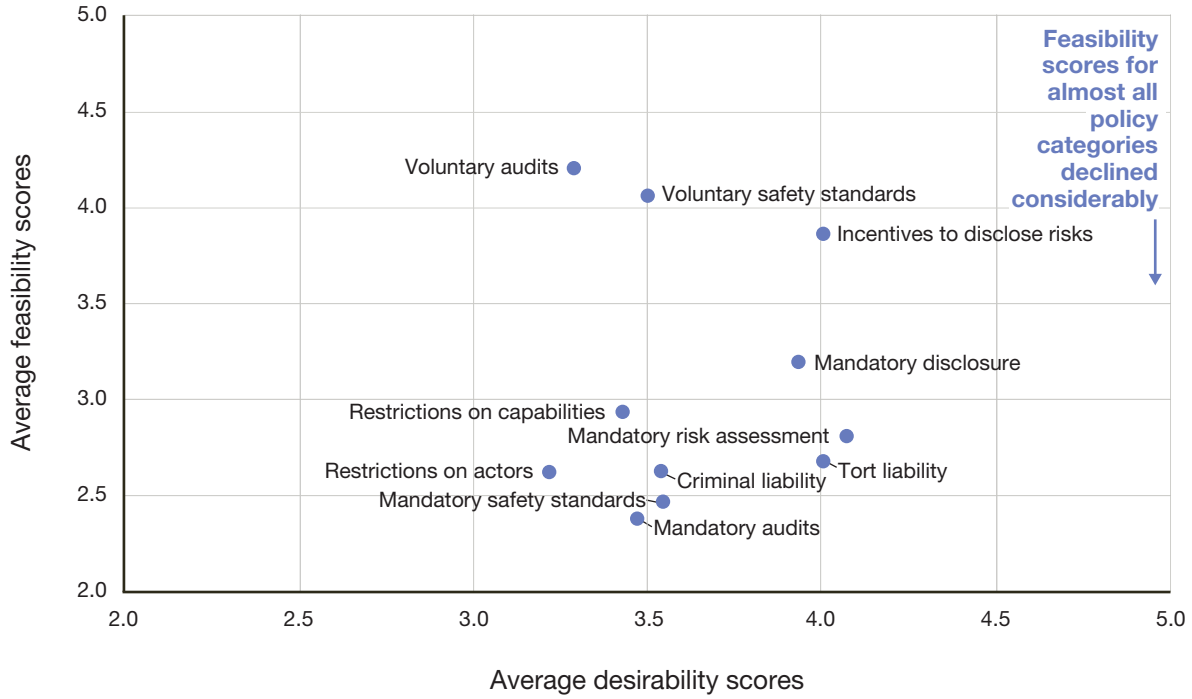
datory audits, with one participant asking, "Do we have even the beginnings of a science that tells us what to audit for? The albeit-brief history of cybersecurity audits . . . suggests they turn into box-checking exercises which emphasize process . . . over product."

By round 3, after participants had a chance to discuss and rescore the policy categories, participants decreased their estimates for almost all policy categories. Most categories were reassessed as less desirable—albeit slightly, with the mean dropping by 0.06 (2 percent). Likewise, most categories were assessed as less effective: The decrease in mean effectiveness scores was marginal for developers (0.16 points, or 5 percent) and somewhat larger for users (0.24 for nonmalicious and 0.18 for malicious, 7 percent for both). The largest change in mean was the decrease in feasibility assessments for almost all policy categories, which decreased by 0.38, or 11 percent. Figure 2 shows the average scores for feasibility and desirability in round 3.

Mandatory categories saw the greatest drops in feasibility assessments: For example, *mandatory risk assessments* dropped by a full point on average. As noted, comments offered by experts for their answers conveyed that some of the shifts occurred after they reviewed and incorporated comments by other participants. For example, one participant explained his downgrade of mandatory safety and security standards by noting that "discussion further convinced me of" the estimate that a "huge amount of energy" would be wasted, "chasing something that is . . . very unlikely to actually be effective."

Much of the shift appears to be due to the change in the priorities of the new administration, with the change in administrations occurring between rounds 1 and 3. In particular, many respondents noted the relatively low appetite for regulation or legislation in the current political environment, which became more evident between the first and second rounds. For example, in regard to mandatory audits, one participant

FIGURE 2
Desirability and Feasibility Scores in Round 3



NOTE: Scores represent the average desirability and feasibility scores across all participants for round 3 (*N* = 16).

noted, "Present changes in the U.S. government suggest that this mode of top-down regulation is less plausible than it seemed a few months ago."

The Most Promising Categories to Shape Incentives for Developers Were Incentives to Find and Disclose Risks and Voluntary Safety Standards

No single set of categories emerged as unequivocally best across all criteria. Indeed, respondents' ratings of the policy and legal categories revealed little overlap among the most effective (for each actor individually), the most feasible, and the most desirable categories, as Table 3 illustrated. How, then, might the most promising legal and policy categories be identified in a way that accounts for all three dimensions? There is no single correct approach to this question, and so we adopted one of several possible ways to distinguish the more promising from the less promising. We opted to weight each dimension equally; to do so, we combined the mean scores for each dimension to produce an overall average rating for each policy for desirability, feasibility, and effectiveness scores separately for developers, nonmalicious users, and malicious users (using data per Table 3), as shown in Table 5.

As shown in Table 5, *incentives to disclose risks* and *voluntary safety standards* consistently ranked as the most promising categories, across all actors.[16] Because AI developers have been the focal actor in policy discussions and research on measures that reduce AI-caused risks, we focus our discussion on the most-promising categories for developers.

[16] We again note that there is only slight separation between scores; nevertheless, there is consistency among the top two policies across all actors.

TABLE 5
More and Less Promising Policy Categories for Each Actor Group

Policy	Developers and Deployers	Nonmalicious Users	Malicious Users
Incentives to disclose risks	**3.81**	**3.65**	**3.44**
Voluntary safety standards	**3.44**	**3.44**	**3.21**
Voluntary audits	3.43	3.33	3.16
Tort liability	3.42	3.41	3.20
Mandatory disclosure	3.42	3.31	2.96
Mandatory risk assessment	3.38	3.36	2.86
Mandatory safety standards	3.24	3.05	2.84
Restrictions on capabilities	3.23	3.26	3.01
Criminal liability	3.22	3.14	3.05
Mandatory audits	3.06	2.77	2.61
Restrictions on actors	2.90	2.92	2.64

NOTE: Values represent average feasibility, desirability, and effectiveness for each actor group for round 3 scores. The top two policy categories are bolded. Although we chose only the top two categories, this selection was arbitrary, and we observed that all remaining policy categories scored only slightly lower than the lower scoring of the top two.

The categories that emerged as more promising generally have some common features. First, the top four categories (*incentives to disclose risks*, *voluntary safety standards*, *voluntary audits*, and, to some extent, *tort liability*) can be implemented without government involvement: AI developers may set up incentives to find and disclose risks (such as bug bounties) and voluntarily commit to safety standards and audits—which, as noted, some AI labs have already done. Even *mandatory disclosure* (the fifth-ranked policy category) would place a lesser burden on government capacity. As Guha et al. (2024) explain, disclosure requirements are commonly deemed to be "easier to implement than other regulatory interventions," in part because they place the "burden of information collection and production [on] the regulated entity" (Guha et al., 2024, pp. 1499, 1500). The same may be said of incentives to find and disclose risks: These measures delegate to various third parties or developers' employees the prerogative to find risks.

Robustness of Findings

To be sure, the difference between the most promising categories and the next few categories is not dramatic, with means differing only by 0.01 in some cases. Thus, we should not overstate the extent to which the top-ranked categories hold greater promise relative to some others, for example. We also examined the median values across all categories and stakeholders and found that results are largely robust to these values (see Appendix C).

There is, however, a more considerable difference between the *more* and *less* promising categories.[17] The least-promising policy categories all consist of more heavy-handed, mandatory measures and generally place a higher burden on government. They require more information, expertise, and/or involvement in implementing safeguards relative to the more promising categories. A government at any level would, for example,

[17] These differences, moreover, are present even if a different approach to identifying the more or less promising categories is adopted, such as median scores, which we discuss in Appendix C.

require a fair amount of expertise and information to identify the appropriate safety and security standards, which undoubtedly would change over time.

Some of the policies within these most promising categories also coincide with those that prior research has identified as the most effective in reducing specific systemic risks and as technically feasible in the European regulatory context. While Uuk et al. (2024, pp. 6–7) do not explicitly separate mandatory from voluntary policy interventions, several of the top eight policies identified as priorities to reduce systemic AI risks are consistent with our findings. For example, safety incident reporting and security information-sharing (which includes disclosure of AI incidents, near misses, and security threats) align with our category of incentives to disclose risk, and input and output filtering and predeployment risk assessments align with our category of voluntary safety standards. Moreover, although not discussed in detail, third-party predeployment model audits and external assessment of testing procedure align with our top category of voluntary audits. These alignments suggest that—notwithstanding different methods of identifying promising or highest-priority measures, different types of risks, and differences in the regulatory environments in the EU and the United States—expert judgments support some of the same policy measures.

Implications: How to Make the Most Promising Policy and Legal Measures More Promising

Even with regard to those categories that emerged as most promising based on the ranking method here, experts surfaced several concerns that may limit the measures' intended effects, trigger unintended effects, and reduce their feasibility. We present those concerns here and discuss how they could be remedied.

We examined expert comments across the three rounds to surface specific concerns that were raised about these more promising categories. We then identified potential remedies to these concerns that were either suggested in our experts' comments or addressed in the literature or existing policy proposals. There are likely ways to mitigate the shortcomings and weaknesses of all categories of legal and policy measures that featured in the elicitation. Here, however, we focus on the two measures that we assessed as more promising—to illustrate that some measures may be crafted in ways that minimize their shortcomings.

Incentives to Find and Disclose Risks

This category comprises various measures that establish incentives—for third parties or for employees of developers—to find and disclose risks that would otherwise remain undiscovered or discovered but undisclosed by developers. This includes whistleblower protections; *bug bounties*, or vulnerability disclosure programs; and crowdsourcing AI evaluation contests and challenges. These measures incentivize many parties to find and disclose risks but also incentivize AI developers to devote more resources toward closing vulnerabilities and remedying them when such vulnerabilities are reported. Overall, this category was assessed to be moderately desirable and feasible (scoring about a 4 on each) and between slightly and moderately effective for developers (with lower effectiveness for users). Measures in this category were overwhelmingly viewed as relatively noncontroversial: "marginal benefit but low cost," as one expert summed up. They were seen as feasible and low cost, in part, because of ample use in other domains and because such measures have already been implemented by major AI developers.[18]

The widely shared assessment that the benefits may be marginal or uncertain was based on several factors. **Technical or scientific infeasibility** remains a central concern, as the current state of the science may

[18] See Appendix B for an overview.

not yet allow for precise definitions or understanding of what constitutes a bug or vulnerability—especially one that could lead to catastrophic harms. Even when such issues are identifiable, they may prove to be minor or unrelated to the most-serious risks. There is also the issue of **increased vulnerabilities**; requiring disclosures may unintentionally expose these weaknesses to malicious actors, and some experts expressed concern that such frameworks might even incentivize the deliberate creation of vulnerabilities in order to benefit from disclosure rewards.

Furthermore, disclosure measures may serve more as symbolic gestures than effective interventions, providing cover for high-risk AI uses while addressing only superficial problems. Another major obstacle is **low government interest**, particularly at the federal level, where many participants viewed the likelihood of meaningful government action as low. This skepticism was compounded by a widespread lack of confidence in federal whistleblower protections. Finally, there is a clear **need for protections for disclosing parties**. Participants emphasized that, beyond inadequate whistleblower statutes, effective disclosure programs must include safeguards to protect individuals from retaliation or adverse consequences when reporting bugs or vulnerabilities.

How might these shortcomings be addressed, especially in the near term? Some shortcomings, by nature, cannot be easily addressed: The need for scientific research cannot easily fill the gap in knowledge in the near term. Other shortcomings of these measures, however, may be mitigated, augmenting their effectiveness, feasibility, and overall desirability, as summarized in Table 6.

As our expert participants observed, while government interest in implementing any policy in this category may be low, private-sector actors, state governments, and nongovernmental actors may be more motivated and well-positioned to do so. Most large AI developers, including OpenAI, Anthropic, Google or Alphabet, and Meta, have already implemented some version of bug bounty programs, although many of these are limited in scope to particular types of risks.[19] Our participants noted that the pressure to adopt or broaden such measures may also come from insurers, which are beginning to cover losses caused by AI (Harris and Heikkilä, 2025; Weil et al., 2024). Whistleblower protections are a common remedy in industries and settings outside AI, including at the state level, and there have been initiatives to adopt these specifically for AI systems (e.g., New Jersey Assembly Resolution 158, 2024; California Senate Bill 53, 2025).

TABLE 6

Shortcomings and Potential Remedies for Incentives to Find and Disclose Risks

Shortcoming	Potential Remedies
Technical or scientific infeasibility	• Research
Low government interest	• AI developer action[a] • Insurer pressure • Nongovernmental participation[a]
Exploitation of vulnerabilities	• Coordinated vulnerability disclosures (CVDs)[a]
Need for greater protection	• Legal safe harbors[a] • Terms of service adjustments[a] • Employment contract adjustments[a]
Symbolic gesture or safety-washing	• Calibrated financial rewards to incentivize finding more-serious vulnerabilities[a]

[a] We consider these remedies to be relatively feasible in the near term.

[19] Anthropic's bug bounty program, for example, is limited to approved applicants and focused on "universal jailbreaks" creating CBRN risks (Anthropic, 2025b); Meta AI's bug bounty scope is focused on "reports that demonstrate integral privacy or security issues associated with Meta's large language models, including being able to leak or extract training data through tactics like model inversion or extraction attacks" (Meta, undated). See also OpenAI (2023).

The potential for the exploitation of vulnerabilities that are publicly disclosed may be mitigated by CVD processes modeled on those that exist in software security, medical devices, and other contexts. Through such a CVD process, a vulnerability would first be disclosed to the vendor and disclosed to the public only after the vendor has had time to address it (e.g., Cattell, Ghosh, and Kaffee, 2024). The use of secure third-party intermediaries to manage bug bounties and CVD processes—as is done for software—also promises to reduce those risks, as well as the related potential perverse incentives to create vulnerabilities in order to reap the rewards of a bug bounty.

As one participant commented, "Researchers and other participants must be granted appropriate protections. Otherwise, participants may be disincentivized from thoroughly probing a system." To adequately incentivize and protect those who find and disclose vulnerabilities, researchers have proposed extending rules for good-faith research that exist in computer security to cover general-purpose AI. Such rules protect those who conduct good-faith research from any "legal or technical retaliation" by the company whose vulnerabilities are discovered (Longpre et al., 2025, p. 7). Such safe-harbor protections for good-faith research may be baked into AI developers' terms of service. Others have suggested that companies can ban nondisparagement or confidentiality clauses from employment contracts (Yeung, 2024).

Finally, the concern that setting up these types of incentives would find only minor risks but could be used as an excuse for AI developers to avoid more-meaningful safeguards is a weighty one. This concern was raised with regard to many of the policy categories considered here and will likely remain present—especially because the science behind identifying bugs or vulnerabilities in AI systems is not yet feasible at scale. Still, there are ways to structure these types of programs to increase the chances that more-serious vulnerabilities or risks would be found: For example, financial incentives may be calibrated to the seriousness of the vulnerability, which may incentivize greater effort (Gal-Or, Hydari, and Telang, 2024).

Voluntary Safety Standards

This category includes efforts to develop and adopt voluntary safety standards for AI development and deployment, including initiatives led by industry, standard-setting bodies, multistakeholder forums, or public-private partnerships that issue nonbinding but publicly visible guidelines. Examples include NIST's AI Risk Management Framework, International Organization for Standardization standards, and Secure by Design pledges. These efforts are generally viewed as highly feasible in the U.S. context. Experts in our elicitation rounds consistently emphasized that voluntary standards are politically expedient, face minimal industry resistance, and are already underway through such entities as NIST and the Cybersecurity and Infrastructure Security Agency.

However, feasibility came at the cost of **uncertain or low effectiveness**, especially in preventing catastrophic-level AI harm. While some experts noted that these frameworks may have downstream accountability effects (e.g., enabling public scrutiny if a firm fails to comply after an incident), many doubted their preventive power. As one participant put it, "Voluntary standards are better than nothing, but I doubt they provide much in terms of effective safety governance." Several experts warned that such standards risk becoming a form of **"responsible AI–washing,"** in which companies accrue reputational benefits without material changes in practice.

Nevertheless, experts also emphasized the value of voluntary standards as a first step, both procedurally and politically. They help form coalitions, catalyze early norm-setting, and provide a lower-stakes proving ground for ideas that may later become regulatory requirements. As one participant remarked, "This may be a useful set of steps to show whether additional action (e.g., legislation or regulation) will be required in the future." Another said, "It's a starting point for deliberation of what real risk mitigation could look like."

While some participants viewed voluntary standards as symbolic, others highlighted that such measures could exert real influence through soft-power mechanisms, including peer pressure, reputational incentives,

procurement requirements, or insurance underwriting. "The public relations impact of not adhering to voluntary best practices could be as damaging as the legal impact of not adhering to mandatory ones," noted one participant. Another said that companies are more likely to voluntarily comply when the costs are low and the benefits (e.g., reputation, investor trust, consumer confidence) are high.

Still, proponents acknowledged that effectiveness depends on who writes the standards and how robust they are. Many warned that industry-led processes, especially without technical consensus or post-incident learning, could produce **"toothless" standards** that merely codify current practices or become outdated as capabilities evolve. Several experts pointed to the cybersecurity field, in which voluntary frameworks, such as the NIST cybersecurity framework, have been influential but were largely reactive and tailored to known threats. With AI, which proactively creates new risks and capabilities, this analogy may falter.

Experts also emphasized the likelihood of **uneven adoption**: Voluntary frameworks tend to reach well-resourced actors first. Larger labs with dedicated policy and safety teams are more likely to adopt best practices voluntarily, whereas smaller firms racing to release competitive models may not follow suit. This dynamic could widen the safety gap and render voluntary norms less impactful at scale. As one participant said, "Voluntary standards won't stop the actors who most need oversight." The shortcomings and potential remedies for voluntary safety standards are summarized in Table 7.

Although experts were divided on the long-term value of voluntary standards, most agreed that they are likely to remain the primary form of AI governance in the near term, especially under conditions of political gridlock or regulatory uncertainty. Several experts suggested that voluntary initiatives could lay the groundwork for future mandates and that their development process could serve as a crucial forum for building technical consensus and institutional capacity. In that light, voluntary standards may not themselves avert catastrophic harm, but they could help define what *responsibility* and *due care* mean in AI development—concepts that courts, regulators, and insurers may increasingly draw on. As one participant said, "Plans are worthless, but planning is invaluable." Voluntary standards, even if imperfect, may become the raw material for the more effective policy frameworks yet to come.

TABLE 7

Shortcomings and Potential Remedies for Voluntary Safety Standards

Shortcoming	Potential Remedies
Low effectiveness in preventing catastrophic harm	Use standards as scaffolding for future regulation[a]
Risk of responsible AI–washing	Strong public commitments and auditing by civil society
Weak enforcement or no consequences for defection	Link to insurance, procurement, or investor requirements[a]
Uneven adoption across developer sizes and resources	Public-sector support or incentives for participation[a]

[a] We consider these remedies to be relatively feasible in the near term.

Conclusion

The rapid progress of AI brings extraordinary capabilities but also the potential for nationally catastrophic harms. The findings of our expert elicitation highlight that, although many governance measures are desirable in principle, their feasibility—particularly within the current U.S. political and regulatory environment—remains a major challenge. Resistance to regulation, preference for market-driven approaches, and the global nature of AI development complicate the creation of robust oversight frameworks.

Even so, this study allows us to differentiate between categories of legal and policy measures judged as more and less promising by 16 participants. Incentives to find and disclose risks and voluntary safety standards emerged as those judged as more promising of the 11 categories we examined.

Experts viewed these measures as more viable than others under conditions as of early 2025, although each has limitations. Importantly, many could be implemented by state governments, industry, and/or nongovernmental actors, reducing reliance on federal capacity. Early adoption by developers—and support from actors, such as insurers—suggests that incremental progress is possible without sweeping legislative reform.

For decisionmakers, these findings suggest three broad opportunities:

- **Foster transparency through incentives.** Low-cost mechanisms, such as well-structured bug bounty programs or rewards for identifying serious vulnerabilities, can improve risk detection and accountability. Secure vulnerability disclosure processes (including staged public release and use of third-party intermediaries) can mitigate exploitation risks. Protective frameworks for those disclosing vulnerabilities—whether at the state level or within companies—can further encourage participation.
- **Establish voluntary standards as a foundation for future mandates.** Voluntary safety frameworks are politically feasible and comparatively easy to implement, especially given the resistance to premature regulation and ongoing uncertainty around AI risk. In the near term, public-private collaborations and industry self-regulation may serve as precursors to more-formal legal requirements. These voluntary measures can lay groundwork for eventual mandates by helping identify which safety norms are operationally viable, socially acceptable, and ready for broader adoption.
- **Use voluntary standards to build consensus and reduce regulatory fragmentation.** Uniform adoption of voluntary frameworks across firms and jurisdictions may reduce the need for immediate federal mandates or hard international law. In settings in which regulation is unlikely in the near term, voluntary measures can still promote baseline safeguards and signal responsible intent. Over time, they may facilitate the transition from fragmented efforts to more-cohesive, enforceable norms.

Across all categories, integrating scientific and empirical evidence into governance design will ultimately be essential (e.g., Bommasani et al., 2024). Any measures need continuous evaluation of their real-world effectiveness, unintended consequences, and economic impact. Understanding the costs for both developers and regulators and identifying potential funding sources or incentives can make promising measures more practical and broadly acceptable.

Rather than wait for perfect consensus and decisive evidence, decisionmakers may wish to pursue incremental, feasible steps that can be implemented now—especially those steps that build transparency, reinforce accountability, and create the conditions for more-robust measures to follow.

APPENDIX A

Pre-Elicitation

The following is the text we provided to a set of RAND experts in the pre-elicitation stage.

We are researching policy interventions designed to address nationally catastrophic harms from AI (see definitions below). Specifically, we are interested in how policies may affect the behaviors or incentives of actors in the causal chain leading to harm, namely

- developers/deployers of AI systems
- malicious users
- nonmalicious users.

We collected policies from research papers, reports, bills, or other U.S. policy that consider alternative approaches to reducing AI harms. The table below [not reproduced here] is a clustered list of these policies. Please review the list (Appendix B) and consider the following questions:

1. Would you add/remove any policies to/from the list? (Note: we are not looking for policies that are so granular as to target a specific topic or application.)
2. Would you change the clustering of the list? If so, how?

As you form your response, consider the following:

- our distinction between mandatory and voluntary enforcement
- how each policy might affect behaviors across the different stakeholders identified above
- the relative effectiveness of the policy in reducing probability of catastrophic harm.

Please document your thoughts either in a reply to this email, or directly in comments in the attached Word document. Once we receive your reply, we will follow up with a conversation to discuss further.

Key Definitions

- AI considered here refers to foundation models, or any model that is trained on broad data using self-supervision at scale, that can be adapted to a wide range of downstream tasks.
- Catastrophic harms are harms that produce "substantial loss of life on large populations or cause tremendous property damage." The thresholds of "substantial" or "tremendous" do not have rigid quantitative definitions but correspond to harms that rise to the level of having national significance. This threshold is considerably lower than "existential harms," which threaten the survival of the species. And it is considerably higher than attritional harms, such as algorithmic bias or discrimination.

- Policies: We do not consider policies such as funding research and development (R&D), or preparedness and response measures to mitigate harms after they occur. Measures such as the latter may well reduce the likelihood that AI causes harm or reduce the magnitude of the harm if it occurs—but not by shaping incentives of actors. We also do not consider ethical or environmental policies that may address discrimination from AI systems, or long-run climate impacts.

APPENDIX B

Description of Categories of Policy and Legal Measures (Policy Sets)

We derived the 11 categories, or policy sets, by examining policies and legal measures discussed in the literature and policy discussions and grouping them based on commonalities. In this appendix, we describe these categories more fully, identify their commonalities, and offer examples. (This material was *not* presented to the participants in the elicitation.)

Mandatory risk assessment and mitigation frameworks and processes. This category covers legal requirements for developers of AI systems to adopt certain internal governance *structures* or *frameworks* or institute *internal processes* aimed at reducing the chances that systems they develop cause catastrophic harms. Mandating certain organizational structures or processes is intended to incentivize developers to better assess the risks posed by their AI systems, enabling developers to identify risks of catastrophic harms in time to mitigate. Although some developers have voluntarily adopted such processes, mandates to this effect are also backed by some types of penalties.[1] Risk assessments are commonly conducted in various industries that have the potential to cause mass or catastrophic harms, such as nuclear power (e.g., Karnofsky, 2024). Proposals specific to AI include predevelopment, predeployment, and continuous assessments. They may be imposed by legislation—at the federal or state level—or issued by a regulatory body.

Mandatory safety and security standards. This category covers a potentially broad set of standards aimed at (1) safe development, deployment, and/or use of AI, to prevent unintended harms, and (2) secure development, deployment, and/or use, to prevent unauthorized, malicious uses. Unlike the first category, which focuses on processes and structures that might detect and mitigate detected risks, these types of measures contain specific standards for the development and/or use of AI that are expected to reduce the likelihood of harms directly. What belongs in this category is likely to evolve over time as technical knowledge grows. Common standards widely thought to augment the safety of AI include the curation of the input training data, reinforcement learning from human feedback, potential curation of requests and responses, and implementations of a variety of "kill switch" or shutdown mechanisms that would stop or pause development when risks emerge (e.g., Gemini Team, 2025; Jindal, 2021; Uuk et al., 2024, p. 11). A frequently proposed security standard pertains to the protection of model weights and algorithms (Nevo et al., 2024). Mandatory standards may be imposed by state or federal governments. They may be imposed by legislation or through regulation by existing regulatory bodies, as well as through hypothetical future regulatory bodies.

Voluntary safety and security standards. This category covers the same substantive set of standards as the previous category but would be adopted voluntarily rather than by state mandate. Such standards may be

[1] For example, OpenAI reports spending "six months on safety research, risk assessment, and iteration prior to launching GPT-4" (OpenAI, 2024a, p. 59). OpenAI states that it follows decision rules based on their risk evaluations; only those that have a score of "medium" risk or below can be deployed, and only those with a score of "high" or below can be developed further (OpenAI, 2024b).

developed and adopted by industry through voluntary commitments or cooperation, such as through individual developer policies or collective AI safety commitments, such as the ones made at Seoul (UK Department for Science, Innovation and Technology, 2025). Voluntary standards may also emerge from AI safety institutes or other standard-setting bodies, as well as hypothetical future self-regulatory organizations. By definition, voluntary standards are not backed by penalties if they are disregarded. However, even voluntary standards may incentivize behavior that reduces the probability of harm—especially when they are publicly announced and accepted as legitimate by the many individuals who play a role in implementing them (e.g., employees of an AI developer).

Mandatory monitoring and audits. This category encompasses various mechanisms for verifying that an AI model is performing as expected and/or a developer is complying with legal requirements or adhering to industry safety standards or policies. Such mechanisms may take the shape of monitoring, audits, or evaluations of AI systems. These may be internal to the developer, with a company evaluating its own models, or external, conducted by governmental or nongovernmental parties (e.g., California Senate Bill 1047, 2024). Monitoring, audits, and evaluations may be mandated by state or federal legislation or imposed by existing or newly formed government regulatory bodies.[2] There may be considerable differences in the strength of the incentives to adhere to standards as between internal and external mechanisms (Bengio et al., 2024). Nonetheless, we group these measures as a single category because of the common underlying logic: Verification provides information to the developer regarding potential risks and to the government body, incentivizing AI development practices that adhere to standards, laws, or policies.

Voluntary monitoring and audits. This category includes all the same verification mechanisms as above but voluntarily adopted. As in the case of other voluntary measures, these may be adopted and overseen by industry through voluntary commitments or cooperation, a standard-setting organization, and/or new self-regulatory organizations.

Restrictions on capabilities or applications. This category includes prohibitions on certain capabilities, applications, or uses of AI systems. Prohibitions may specify actions or capabilities that actors cannot do, have, or develop or identify a general class of capabilities that cannot be developed. For example, Article 5, paragraph 1(a), of the EU AI Act prohibits AI models that it classifies as "unacceptable risk," such as certain systems that deploy "purposefully manipulative or deceptive techniques" (EU, 2024). The world's leading AI scientists have identified other redlines, such as prohibitions on autonomous replication or improvement, power-seeking, deception, assistance for weapon development, and conducting cyberattacks (Bengio, 2024). Such restrictions may be imposed by state or federal governments or voluntarily adopted by industry (e.g., OpenAI, 2025).[3] Indeed, some U.S. states have adopted or proposed restrictions not on capabilities but on specific AI models, at least for some purposes—such as bans or proposed bans on using Chinese AI models (e.g., DeepSeek) on state government–issued devices (Freedman, 2025). AI developers also commit to restricting their systems in certain ways, with leading AI companies committing not to deploy models "capable of causing catastrophic harm" (Anthropic, 2025a, p. 1) or those "that create new risks of severe harm" (OpenAI, 2025, p. 1) until they have built safeguards to minimize those risks.

Restrictions on actors. Another category of restrictions entails limiting or regulating who or what actors are allowed to develop, deploy, and/or use AI systems. Such restrictions may be accomplished through various measures. For example, licensing for developers or deployers of frontier or high-risk AI models has been

[2] The EU AI Act, for example, requires evaluations of general purpose AI models with systemic risk and third-party "conformity assessments" (audits) of "high-risk" AI systems before they are placed on the market or deployed (EU, 2024).

[3] We did not distinguish between mandatory and voluntary measures for this category, largely because of the need to maintain a tractable number of categories.

proposed by both legislators and AI companies.[4] Restrictions on who may develop specific AI systems may be accomplished through establishing regulatory sandboxes for particular developers to do so in a controlled environment. Restricting the set of actors who can develop sophisticated AI models may also be accomplished through export controls for foreign developers or other restrictions of access to compute (Matias, 2023). Restrictions may also be applied to users through Know Your Customer screening mandates.

Incentives to find and disclose risks. This category includes measures that create incentives—for third parties, users, and employees of developers—to increase chances that risky or unlawful practices are exposed. These measures incentivize many parties to find and disclose risks but also incentivize AI developers to devote more resources toward patching vulnerabilities, which would otherwise not be disclosed, and remedy them when such vulnerabilities are reported. Incentives to find and disclose risks may be created by state or federal governments or voluntarily created by industry through such measures as whistleblower protections, bug bounties or vulnerability disclosure programs, or crowdsourcing AI evaluation contests and challenges. Such measures have been commonly proposed by researchers, nongovernmental organizations, and state governments (e.g., Barrett et al., 2023; California Senate Bill 1047, 2024). Whistleblower protections are a common remedy in industries and settings outside AI, and there have been initiatives to adopt these specifically for AI systems (e.g., New Jersey Assembly Resolution 158, 2024). Most large AI developers, including OpenAI, Anthropic, Google or Alphabet, and Meta, have already implemented some version of bug bounty programs, although many of these are focused or limited in scope to particular types of risks or limited to only some "bounty hunters."

Mandatory disclosure. This category consists of varied requirements for AI system developers or deployers to disclose or share information about AI risks, vulnerabilities, or aspect of its performance, safety, training data, design, or downstream applications. Disclosure laws are a staple tool of regulation across a multitude of sectors and are seen as a valuable way to detect and identify potential harms and inform responses (Guha et al., 2024). These laws may include a requirement to create and publicly disclose an inventory of AI use cases (e.g., Executive Order 13960, 2020); to register companies that develop sophisticated general-purpose AI models, or models used in high-risk situations; to disclose risks or impact assessments; or to report security and safety incidents. The mandated disclosures may be to the public or only to the government (Hadfield, Cuéllar, and O'Reilly, 2023). While the specific information to be disclosed and how may vary, the common denominator for measures in this category is that they incentivize a degree of transparency, potentially reducing the likelihood of a development of a dangerous AI system or capability in secret. Disclosures may be mandated by state or federal government and could be to the government, the public, or other AI developers.

Tort liability. Unlike the prior categories, tort liability does not require novel interventions to apply to potential cases of AI harms. Should such harms come about, AI developers—and even users—may face tort liability for their role in causing these harms. Compensatory and punitive monetary damages may be imposed on liable parties. Although tort liability is imposed ex post, its prospect also incentivizes due care by parties facing liability risks ex ante, thereby potentially contributing to practices that reduce risks of catastrophic harms. Tort liability relies on private individuals (victims) or government agents to bring legal actions against parties responsible for the harm. Claims may be brought under existing tort law, in which various tort theories may be invoked to hold developers or deployers and users liable for harms that occur on

[4] For example, OpenAI CEO Sam Altman (2023) proposed licensing as part of a comprehensive regulatory framework in his Senate testimony. Anderljung et al. (2023) suggest that if AI models "pose risks to public safety above a high threshold of severity," frontier AI developers should obtain a license "to widely deploy" the frontier AI model after demonstrating compliance with safety standards. The Blumenthal-Hawley Bipartisan Framework for U.S. AI Act proposes to establish an "independent oversight body" to grant licenses to companies that seek to deploy "sophisticated general purpose A.I. models" and AI used in "high-risk situations" (Blumenthal and Hawley, 2023).

the basis of negligence, recklessness, or intent (Ramakrishnan, Smith, and Downey, 2024). State or federal government may also adopt new laws defining tort liability for AI catastrophic harm, as some have argued should be done (Weil, 2024; Weil et al., 2024). For example, strict liability may be imposed in certain situations by legislation (such as for an abnormally dangerous activity).

Criminal liability. Criminal liability empowers government agents to bring criminal charges against parties responsible for the harm (developers, deployers, or users). In addition to monetary sanctions (as in tort liability), the punishment for those found guilty may include deprivation of liberty and losses of civil rights. Charges may be brought under existing laws. Any number of existing general crimes may apply, depending on the harm caused, with different act and intent requirements—for example, reckless endangerment, intentionally or knowingly causing a catastrophe (in some jurisdictions), or manslaughter. Alternatively, state or federal government may define new criminal laws specific to AI-caused harms.

APPENDIX C

Median Values of Participant Responses

In Table C.1, we show the median values of participant responses for desirability, feasibility, and effectiveness for all stakeholders.

The averages of the median scores for developers and deployers, nonmalicious users, and malicious users are shown in Table C.2. Among the top-ranked policies, *incentives to disclose risks* and *voluntary safety standards* are the two highest-ranked policies also from the mean analysis. Because of the range compression of the five-point Likert scale, however, there is less variation among top scores, driving two additional policies to be ranked in second place.

TABLE C.1
Median Values of Participant Responses for Desirability, Feasibility, and Effectiveness

Policy	Desirability	Feasibility	Effectiveness for Developers	Effectiveness for Nonmalicious Users	Effectiveness for Malicious Users
Incentives to disclose risks	4	4	4	3	2
Voluntary safety standards	4	4	3	3	2
Mandatory disclosure	4	4	3	3	2
Tort liability	4	3	4	4	3
Mandatory safety standards	4	2	4	3	2.5
Mandatory audits	4	2	4	2	2
Voluntary audits	3	4	3	2	2
Criminal liability	4	2	4	3	3
Mandatory risk assessment	4	2.5	3	3	2
Restrictions on capabilities	3	3	3	3	2.5
Restrictions on actors	3	2.5	2.5	3	2

TABLE C.2

Average of Median Scores for All Actors

Policy	Average Score for Developers	Average Score for Nonmalicious Users	Average Score for Malicious Users
Incentives to disclose risks	4.00	3.67	3.33
Voluntary safety standards	3.67	3.67	3.33
Mandatory disclosure	3.67	3.67	3.33
Tort liability	3.67	3.67	3.33
Mandatory safety standards	3.33	3.00	2.83
Mandatory audits	3.33	2.67	2.67
Voluntary audits	3.33	3.00	3.00
Criminal liability	3.33	3.00	3.00
Mandatory risk assessment	3.17	3.17	2.83
Restrictions on capabilities	3.00	3.00	2.83
Restrictions on actors	2.67	2.83	2.50

Abbreviations

AI	artificial intelligence
CBRN	chemical, biological, radiological, and nuclear
CVD	coordinated vulnerability disclosure
EU	European Union
HAI	Stanford Institute for Human-Centered Artificial Intelligence
NIST	National Institute of Standards and Technology
R&D	research and development
UK	United Kingdom

References

Altman, Samuel, "Oversight of A.I.: Rules for Artificial Intelligence," testimony before the Subcommittee on Privacy, Technology, and the Law of the Senate Judiciary Committee, 118th Congress, first session, U.S. Government Publishing Office, May 16, 2023.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv, arXiv:2307.03718v4, November 7, 2023.

Anthropic, *Responsible Scaling Policy*, version 2.1, March 31, 2025a.

Anthropic, "Testing Our Safety Defenses with a New Bug Bounty Program," webpage, May 14, 2025b. As of February 26, 2026:
https://www.anthropic.com/news/testing-our-safety-defenses-with-a-new-bug-bounty-program

Barrett, Anthony M., Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, and Krystal Jackson, *AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models*, version 1.0, UC Berkeley Center for Long-Term Cybersecurity, November 2, 2023.

Beckers, Anna, and Gunther Teubner, *Three Liability Regimes for Artificial Intelligence: Algorithmic Actants, Hybrids, Crowds*, Hart Publishing, 2002.

Bengio, Yoshua, "Government Interventions to Avert Future Catastrophic AI Risks," *Harvard Data Science Review*, Special Issue 5, June 2024.

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al., "Managing Extreme AI Risks amid Rapid Progress," *Science*, Vol. 384, No. 6698, May 20, 2024.

Blumenthal, Richard, and Josh Hawley, "Bipartisan Framework for U.S. AI Act," 2023. As of March 9, 2026:
https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf

Bommasani, Rishi, Sanjeev Arora, Yejin Choi, Daniel E. Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, Fei-Fei Li, Arvind Narayanan, Alondra Nelson, et al., "A Path for Science- and Evidence-Based AI Policy," Understanding AI Safety, 2024.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al., "On the Opportunities and Risks of Foundation Models," arXiv, arXiv:2108.07258v3, July 12, 2022.

Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," arXiv, arXiv:1802.07228v2, December 1, 2024.

California Senate Bill 53, Artificial Intelligence Models: Large Developers, September 29, 2025.

California Senate Bill 1047, Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, September 3, 2024.

Carlsmith, Joseph, "Is Power-Seeking AI an Existential Risk?" arXiv, arXiv:2206.13353v2, August 13, 2024.

Cattell, Sven, Avijit Ghosh, and Lucie-Aimée Kaffee, "Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities," arXiv, arXiv:2402.07039v3, July 26, 2024.

Center for AI Safety, homepage, undated. As of March 2, 2026:
https://safe.ai/ai-risk

Centre for the Study of Existential Risk, homepage, undated. As of March 2, 2026:
https://www.cser.ac.uk

Cohen, Michael K., Noam Kolt, Yoshua Bengio, Gillian K. Hadfield, and Stuart Russell, "Regulating Advanced Artificial Agents," *Science*, Vol. 384, No. 6691, April 4, 2024.

EU—*See* European Union.

European Parliament, "EU AI Act: First Regulation on Artificial Intelligence," last updated February 19, 2025.

European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonized Rules on Artificial Intelligence and Amending Various Regulations and Directives (Artificial Intelligence Act), 2024.

Executive Order 13960, "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," Executive Office of the President, December 3, 2020.

Freedman, Linn F., "Three States Ban DeepSeek Use on State Devices and Networks," *National Law Journal*, February 13, 2025.

Future of Humanity Institute, homepage, undated. As of March 2, 2026:
https://www.futureofhumanityinstitute.org

Future of Life Institute, homepage, undated. As of March 2, 2026:
https://futureoflife.org

Gal-Or, Esther, Muhammad Zia Hydari, and Rahul Telang, "Merchants of Vulnerabilities: How Bug Bounty Programs Benefit Software Vendors," arXiv, arXiv:2404.17497, April 26, 2024.

Gemini Team, "Gemini: A Family of Highly Capable Multimodal Models," arXiv, arXiv:2312.11805v5, May 9, 2025.

Guha, Neel, Christie M. Lawrence, Lindsey A. Gailmard, Kit T. Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, et al., "AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing," *George Washington Law Review*, Vol. 92, No. 6, December 2024.

Hadfield, Gillian, Mariano-Florentino (Tino) Cuéllar, and Tim O'Reilly, "It's Time to Create a National Registry for Large AI Models," Carnegie Endowment for International Peace, July 12, 2023.

HAI—*See* Stanford Institute for Human-Centered Artificial Intelligence.

Harris, Jeremie, Edouard Harris, and Mark Beall, *Survey of AI Technologies and AI R&D Trajectories*, Gladstone AI, November 3, 2023.

Harris, Lee, and Melissa Heikkilä, "Insurers Launch Cover for Losses Caused by AI Chatbot Errors," *Financial Times*, May 11, 2025.

Jindal, Sonam, "Responsible Sourcing of Data Enrichment Services," *Partnership on AI* blog, June 16, 2021. As of February 26, 2026:
https://partnershiponai.org/responsible-sourcing-considerations

Karnofsky, Holden, "Developing AI Risk Management with the Same Ambition and Urgency as AI Products," Carnegie Endowment for International Peace, December 16, 2024.

Khodyakov, Dmitry, Sean Grant, Jack Kroger, and Melissa Bauman, *RAND Methodological Guidance for Conducting and Critically Appraising Delphi Panels*, RAND Corporation, TL-A3082-1, 2023. As of May 8, 2025:
https://www.rand.org/pubs/tools/TLA3082-1.html

Longpre, Shayne, Kevin Klyman, Ruth E. Appel, Sayash Kapoor, Rishi Bommasani, Michelle Sahar, Sean McGregor, Avijit Ghosh, Borhane Blili-Hamelin, Nathan Butters, et al., "In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI," arXiv, arXiv:2503.16861v2, March 25, 2025.

Matias, Jeanina, "What Is a Foundation Model? An Explainer for Non-Experts," Stanford Institute for Human-Centered Artificial Intelligence, May 10, 2023.

Meta, "Meta Bug Bounty Scope," webpage, undated. As of February 26, 2026:
https://bugbounty.meta.com/scope

National Institute of Standards and Technology, *Standards for Security Categorization of Federal Information and Information Systems*, U.S. Department of Commerce, Federal Information Processing Standards Publication 199, February 2004.

National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, U.S. Department of Commerce, NIST AI 100-1, January 2023.

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND Corporation, RR-A2849-1, 2024. As of February 26, 2026:
https://www.rand.org/pubs/research_reports/RRA2849-1.html

New Jersey Assembly Resolution 158, 221st Leg., September 23, 2024.

New York State Assembly Bill 6453, Responsible AI Safety and Education Act, 2025–2026 Leg., Reg. Sess., March 5, 2025.

NIST—*See* National Institute of Standards and Technology.

Norman, Geoff, "Likert Scales, Levels of Measurement and the 'Laws' of Statistics," *Advances in Health Sciences Education*, Vol. 15, No. 1, March 2010.

OpenAI, "Announcing OpenAI's Bug Bounty Program," webpage, April 11, 2023. As of February 26, 2026:
https://openai.com/index/bug-bounty-program

OpenAI, "GPT-4 Technical Report," arXiv, arXiv:2303.08774v6, March 4, 2024a.

OpenAI, "OpenAI o1 System Card," webpage, last updated December 5, 2024b. As of February 26, 2026:
https://openai.com/index/openai-o1-system-card/

OpenAI, *Preparedness Framework*, version 2, April 15, 2025.

Pan, Alexander, Kush Bhatia, and Jacob Steinhardt, "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models," arXiv, arXiv:2201.03544v2, February 14, 2022.

Ramakrishnan, Ketan, Gregory Smith, and Conor Downey, *U.S. Tort Liability for Large-Scale Artificial Intelligence Damages: A Primer for Developers and Policymakers*, RAND Corporation, RR-A3084-1, 2024. As of February 25, 2026:
https://www.rand.org/pubs/research_reports/RRA3084-1.html

Schuett, Jonas, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel, "Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion," arXiv, arXiv:2305.07153, May 11, 2023.

Shepardson, David, and Anna Tong, "California Governor Vetoes Contentious AI Safety Bill," Reuters, September 30, 2024.

Stanford Institute for Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report*, Stanford University, 2025.

UK Department for Science, Innovation and Technology—*See* United Kingdom Department for Science, Innovation and Technology.

United Kingdom Department for Science, Innovation and Technology, "Frontier AI Safety Commitments, AI Seoul Summit 2024," February 7, 2025.

Uuk, Risto, Carlos Ignacio Gutierrez, Daniel Guppy, Lode Lauwaert, Atoosa Kasirzadeh, Lucia Velasco, Peter Slattery, and Carina Prunkl, "A Taxonomy of Systemic Risks from General-Purpose AI," arXiv, arXiv:2412.07780, November 24, 2024.

Viscusi, W. Kip, and Richard Zeckhauser, "Addressing Catastrophic Risks: Disparate Anatomies Require Tailored Therapies," in Cary Coglianese, ed., *Regulatory Breakdown: The Crisis of Confidence in U.S. Regulation*, University of Pennsylvania Press, 2012.

Weil, Gabriel, "Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence," Touro University, Jacob D. Fuchsberg Law Center, 2024.

Weil, Gabriel, Matteo Pistillo, Suzanne Van Arsdale, Junichi Ikegami, Kensuke Onuma, Megumi Okawa, and Michael A. Osborne, *Insuring Emerging Risks from AI*, AI Governance Initiative, Oxford Martin School, University of Oxford, November 14, 2024.

White House, *Winning the Race: America's AI Action Plan*, Executive Office of the President, July 2025.

Yeung, Douglas, "AI Companies Say Safety Is a Priority. It's Not," *San Francisco Chronicle*, July 9, 2024.

About the Authors

Sasha Romanosky is a senior policy researcher at RAND, where he studies topics on the economics of security and privacy, AI, cyber insurance, cybercrime, and law and economics. He holds a Ph.D. in public policy.

Elina Treyger is a senior political scientist at RAND working at the intersection of homeland security, defense, and international affairs, including topics related to strategic competition, hybrid warfare, and AI. She holds a Ph.D. in political science.

Elie Alhajjar is a senior information scientist at RAND leading projects at the intersection of AI, cybersecurity, and government and efforts around the commercial space sector, Army acquisition, adversarial machine learning, cyber risk, and countering weapons of mass destruction. He holds a Ph.D. in applied mathematics.