Turning Vision into Action: Implementing the Senate AI Roadmap

17th June 2024

US Policy Team
policy@futureoflife.org

Contents

Executive Summary

On May 15, 2024, the Senate AI Working Group released "Driving U.S. Innovation in Artificial Intelligence: A Roadmap for Artificial Intelligence Policy in the United States Senate," which synthesized the findings from the Senate AI Insight Forums into a set of recommendations for Senate action moving forward. The Senate's sustained efforts to identify and remain abreast of the key issues raised by this rapidly evolving technology are commendable, and the Roadmap demonstrates a remarkable grasp of the critical questions Congress must grapple with as AI matures and permeates our everyday lives.

The need for regulation of the highest-risk AI systems is urgent. The pace of AI advancement has been frenetic, with Big Tech locked in an out-of-control race to develop increasingly powerful, and increasingly risky, AI systems. Given the more deliberate pace of the legislative process, we remain concerned that the Roadmap's deference to committees for the development of policy frameworks could delay the passage of substantive legislation until it is too late for effective policy intervention.

To expedite the process of enacting meaningful regulation of AI, we offer the following actionable recommendations for such policy frameworks that can form the basis of legislation to reduce risks, foster innovation, secure wellbeing, and strengthen global leadership.

AGI and Testing of Advanced General-Purpose AI Systems

- Adopt a stratified, capabilities-based oversight framework including pre-deployment assessment, auditing, and licensure, and post-deployment monitoring for the most advanced general-purpose AI systems (GPAIS), with the burden of proving the system's suitability for release resting on the developer of the system.
- Require audits evaluating the safety and security of advanced GPAIS to be conducted by independent, objective actors either within the government or accredited by the government, with the authority to prohibit the release of an advanced AI system if they identify that it poses a significant unresolved risk to the safety, security, or wellbeing of the American public.
- Establish regulatory thresholds that are inclusive of, but not limited to, training compute to ensure that current and future systems of concern remain in scope, with each threshold independently sufficient to qualify a system as subject to additional scrutiny.
- Establish a centralized federal authority responsible for monitoring, evaluating, and regulating advanced GPAIS, and for advising other agencies on activities related to AI within their respective jurisdictions.
- Augment whistleblower protections to cover reporting on unsafe practices in the development or planned deployment of AI systems.

Liability

- Specify strict liability ("abnormally dangerous activity") for the development of the most advanced GPAIS, which pose risks that cannot be anticipated and therefore cannot be eliminated through reasonable care.
- Apply a rebuttable presumption of negligence for harms caused by advanced GPAIS that do not meet the threshold to be subject to strict liability but do meet a lower threshold of capability.
- Subject domain-specific AI systems to the legal standards applicable in that domain.

- Apply joint and several liability for harms caused by advanced GPAIS.
- Clarify that Section 230 does not shield AI providers from liability for harms resulting from their systems.

AI and National Security

- Establish an Information Sharing and Analysis Center and require AI developers to share documentation with the ISAC pertaining to the development and deployment lifecycle.
- Require the most powerful AI systems and those that could pose CBRN threats to be tested in an AIxBio sandbox as proposed in the 2025 NDAA draft.
- Prohibit training models on the most dangerous dual-use research of concern and restrict the use of dual-use research of concern for training narrow AI systems.
- Invest in core CBRN defense strategies such as personal protective equipment, novel medical countermeasures, and ultraviolet-c technologies.

Compute Security and Export Controls

- Support the passage of the ENFORCE Act (H.R. 8315) with clarifications to avoid loopholes from open publication of model weights.
- Require the use of chips with secure hardware for the training of advanced AI systems and to receive licensure for export of high-end chips to restricted countries.

Autonomous Weapons Systems and Military Integration of AI

- Mandate that nuclear launch systems remain independent from CJADC2 capabilities
- Require DOD to establish boards comprised of AI ethics officers across offices involved in the production, procurement, development, and deployment of military AI systems.
- Task CDAO with establishing clear protocols for measuring the accuracy of AI systems integrated into defense operations and prohibit integration of systems that do not meet the threshold of "five-digit accuracy."
- Codify DOD Directive 3000.09 and raise required human involvement in military use of AI from "appropriate levels of human judgment" to "meaningful human control"; require CDAO to file a report establishing concrete guidance for meaningful human control in practice.
- Invest in the development of non-kinetic counter-autonomous weapons systems.

Open-Source AI

- Require advanced AI systems with open model weights to undergo thorough testing and evaluation in secure environments appropriate to their level of risk.
- Hold developers legally responsible for performing all reasonable measures to prevent their models from being retrained to enable illegal activities, and for harms resulting from their failure to do so.
- Pursue "public options" for AI such as the National AI Research Resource (NAIRR) to democratize AI development and combat concentration of power.

Supporting AI Innovation

- Allocate R&D funding to BIS for the development of on-chip hardware governance solutions, and for the implementation of those solutions.
- Expand NAIRR public compute programs to include funding directed toward the development of secure testing and usage infrastructure for academics, researchers, and members of civil society.
- Ensure that R&D funding allocated towards improving interagency coordination at the intersection of AI and critical infrastructure includes funding requirements to safety and security research.

Combatting Deepfakes

- Create civil and/or criminal liability mechanisms to hold developers and providers accountable for harms resulting from deepfakes.
- Ensure users accessing models to produce and share deepfakes are subject to civil and/or criminal liability.
- Place a responsibility on compute providers to revoke access to their services when they have knowledge that their services are being used to create harmful deepfakes, or to host models that facilitate the creation of harmful deepfakes.
- Support the passage of proposed bills including the NO FAKES Act, with some modifications to clarify the liability of service providers such as model developers.

Provenance and Watermarking

- Require model developers and providers to integrate provenance tracking capabilities into their systems.
- Require model developers and providers to make content provenance information as difficult to bypass or remove as possible, taking into account the current state of science.
- Support the passage of bills like the AI Labeling Act, which mandate clear and permanent notices on AI-generated content that identify the content as AI-produced and specify the tool used along with the creation date.

Introduction

In May 2024, the Bipartisan Senate AI Working Group, spearheaded by Majority Leader Schumer, Sen. Rounds, Sen. Heinrich, and Sen. Young, released a "roadmap for artificial intelligence policy in the United States Senate" entitled "Driving U.S. Innovation in Artificial Intelligence." The Roadmap is a significant achievement in bipartisan consensus, and thoughtfully identifies the diversity of potential avenues AI presents for both flourishing and catastrophe. Drawing on the input of experts at the Senate AI Insight Forums and beyond, the Roadmap includes several promising recommendations for the Senate's path forward.

At the same time, the Roadmap lacks the sense of urgency for congressional action we see as critical to ensuring AI is a net benefit for the wellbeing of the American public, rather than a source of unfettered risk. The pace of advancement in the field of AI has accelerated faster than even leading experts had anticipated, with competitive pressures and profit incentives driving Big Tech companies to race haphazardly toward creating more powerful, and consequently less controllable, systems by the month. A byproduct of this race is the relegation of safety and security to secondary concerns for these developers.

The speed with which this technology continues to evolve and integrate stands in stark contrast to the typical, more deliberate pace of government. This mismatch raises a risk that requisite government oversight will not be implemented quickly enough to steer AI development and adoption in a more responsible direction. Realization of this risk would likely result in a broad array of significant harms, from systematic discrimination against disadvantaged communities to the deliberate or accidental failure of critical infrastructure, that could otherwise be avoided. The social and economic permeation of AI could also render future regulation nearly impossible without disrupting and potentially destabilizing the US's socioeconomic fabric – as we have seen with social media, reactive regulation of emerging technology raises significant obstacles where proactive regulation would not, and pervasive harm is often the result. In other words, the time to establish meaningful regulation and oversight of advanced AI is now.

With this in mind, we commend the Senate AI Working Group for acting swiftly to efficiently bring the Senate up to speed on this rapidly evolving technology through the Senate AI Insight Forums and other briefings. However, we are concerned that, in most cases, the Roadmap encourages committees to undertake additional consideration toward developing frameworks from which legislation could then be derived, rather than contributing to those actionable frameworks directly. We recognize that deference to committees of relevant jurisdiction is not unusual, but fear that this process will imprudently delay the implementation of AI governance, particularly given the November election's potential to disrupt legislative priorities and personnel.

To streamline congressional action, we offer concrete recommendations for establishing legislative frameworks across a range of issues raised in the Roadmap. Rather than building the necessary frameworks from the ground up, our hope is that the analyses and recommendations included herein will provide actionable guidance for relevant committees and interested members that would reduce risks from advanced AI, improve US innovation, wellbeing, and global leadership, and meet the urgency of the moment.

AGI and Testing of Advanced General-Purpose AI Systems

We applaud the AI Working Group for recognizing the unpredictability and risk associated with the development of increasingly advanced general-purpose AI systems (GPAIS). The Roadmap notes "the significant level of uncertainty and unknowns associated with general purpose AI systems achieving AGI." We caution, however, against the inclination that the uncertainty and risks from AGI manifest only beyond a defining, rigid threshold, and emphasize that these systems exist on a spectrum of capability that correlates with risk and uncertainty. Unpredictability and risks have already been observed in the current state-of-the-art, which most experts categorize as sub-AGI, and are expected to increase in successive generations of more advanced systems, even as new risks emerge.

While the Roadmap encourages relevant committees to identify and address gaps in the application of existing law to AI systems within their jurisdiction, the general capabilities of these systems make it particularly challenging to identify appropriate committees of jurisdiction as well as existing legal frameworks that may apply. This challenge was a major impetus for establishing the AI Working Group — as the Roadmap notes in the Introduction, "the AI Working Group's objective has been to complement the traditional congressional committee-driven policy process, considering that this broad technology does not neatly fall into the jurisdiction of any single committee."

Rather than a general approach to regulating the technology, the Roadmap suggests addressing the broad scope of AI risk through use case-based requirements on high-risk uses of AI. This approach may indeed be appropriate for most AI systems, which are designed to perform a particular function and operate exclusively within a specific domain. For instance, while some tweaks may be necessary, AI systems designed exclusively for financial evaluation and prediction can reasonably be overseen by existing bodies and frameworks for financial oversight. We are also pleased by the AI Working Group's acknowledgement that some more egregious uses of AI should be categorically banned – the Roadmap specifically recommends a prohibition on the use of AI for social scoring, and encourages committees to "review whether other potential uses for AI should be either extremely limited or banned."

That said, a use case-based approach is not sufficient for today's most advanced GPAIS, which can effectively perform a wide range of tasks, including some for which they were not specifically designed, and can be utilized across distinct domains and jurisdictions. If the same system is routinely deployed in educational, medical, financial, military, and industrial contexts but is specialized for none of them, the governing laws, standards, and authorities applicable to that system cannot be easily discerned, complicating compliance with existing law and rendering regulatory oversight cumbersome and inefficient.

Consistent with this, the Roadmap asks committees to "consider a capabilities-based AI risk regime that takes into consideration short-, medium-, and long-term risks, with the recognition that model capabilities and testing and evaluation capabilities will change and grow over time." In the case of GPAIS, such a regime would categorically include particular scrutiny for the most capable GPAIS, rather than distinguishing them based on the putative use-case.

Metrics

Our ability to preemptively assess the risks and capabilities of a system is currently limited. As the Roadmap prudently notes, "[a]s our understanding of AI risks further develops, we may discover better risk-management regimes or mechanisms. Where testing and evaluation are insufficient to directly measure capabilities, the AI Working Group encourages the relevant committees to explore proxy metrics that may be used in the interim."

While substantial public and private effort is being invested in the development of reliable benchmarks for assessment of capabilities and associated risk, the field has not yet fully matured. Though some metrics exist for testing the capabilities of models at various cognitive tasks, no established benchmarks exist for determining their capacity for hazardous behavior without extensive testing across multiple metrics.

The number of floating-point operations (FLOPs) are a measure of computation, and, in the context of AI, reflect the amount of computational resources ("compute") used to train an AI model. Thus far, the amount of compute used to effectively train an AI model scales remarkably well with the general capabilities of that model. The recent flurry of advancement in AI capabilities over the past few years has been primarily driven by innovations in high-performance computing infrastructure that allow for leveraging more training data and computational power, rather than from major innovations in model design. The resulting models have demonstrated capabilities highly consistent with predictions based on the amount of computing power used in training, and capabilities have in turn consistently correlated with identifiable risks.

While it is not clear whether this trend will continue, training compute has so far been the objective, quantitative measurement that best predicts the capabilities of a model prior to testing. In a capabilities-based regulatory framework, such a quantitative threshold is essential for initially delineating models subject to certain requirements from those that are not. That said, using a single proxy metric as the threshold creates the risk of failing to identify potentially high-risk models as advances in technology and efficiency are made, and of gamesmanship to avoid regulatory oversight.

Recommendations

1. **Congress should implement a stratified, capabilities-based oversight framework for the most advanced GPAIS to complement use case-dependent regulatory mechanisms for domain-specific AI systems.** Such a framework, through pre-deployment assessment, auditing, and licensure, and post-deployment monitoring, could conceivably mitigate risks from these systems regardless of whether they meet the relatively arbitrary threshold of AGI. While establishing a consensus definition of AGI is a worthwhile objective, it should not be considered prerequisite to developing a policy framework designed to mitigate risks from advanced GPAIS.
2. **Regulatory oversight of advanced GPAIS should employ the precautionary principle, placing the burden of proving the safety, security, and net public benefit of the system, and therefore its suitability for release, on the developer, and should prohibit the release of the system if it does not demonstrate such suitability.** The framework should impose the most stringent requirements on the most advanced systems, with fewer regulatory requirements for less capable systems, in order to avoid unnecessary red tape and minimize the burden on smaller AI developers who lack the financial means to train the most powerful systems regardless.
3. **Audits and assessments should be conducted by independent, objective third-parties who lack financial and other conflicts of interest.** These auditors could either be employed by the government or accredited by the government to ensure they are bound by standards of practice. For less powerful and lower risk systems, some assessments could be conducted in-house to reduce regulatory burden, but verifying the safety of the highest-risk systems should under no circumstances rely on self-governance by profit-motivated companies.
4. **Legislation governing advanced GPAIS should adopt regulatory thresholds that are inclusive of,**

but not limited to, training compute to ensure that current and future systems of concern remain in scope. Critically, these thresholds should each independently be sufficient to qualify a system as subject to additional scrutiny, such that exceeding, e.g. 10^{25} FLOPs in training compute OR 100 billion parameters OR 2 trillion tokens of training data OR exceeding a particular score on a specified capabilities benchmark, risk assessment benchmark, risk assessment rubric, etc.,[1] would require a model to undergo independent auditing and receive a license for distribution. This accounts for potential blindspots resulting from the use of proxy metrics, and allows flexibility for expanding the threshold qualifications as new benchmarks become available.

5. **Congress should establish a centralized federal authority responsible for monitoring, evaluating, and regulating GPAIS** due to their multi-jurisdictional nature, and for advising other agencies on activities related to AI within their respective jurisdictions. This type of "hub and spoke" model for an agency has been effectively implemented for the Cybersecurity and Infrastructure Security Agency (CISA), and would be most appropriate for the efficient and informed regulation of AI. Such an agency could also lead response coordination in the event of an emergency caused by an AI system. Notably, CISA began as a directorate within the Department of Homeland Security (National Protection and Programs Directorate), but was granted additional operational independence thereafter. A similar model for an AI agency could mitigate the logistical and administrative strain that could delay establishment of a brand new agency, with the Department of Energy or the Department of Commerce serving as the hub for incubating the new oversight body.
6. **Whistleblower protections should be augmented to cover reporting on unsafe practices in development and/or planned deployment of AI systems.** It is not presently clear whether existing whistleblower protections for consumer product safety would be applicable in these circumstances; as such, new regulations may be necessary to encourage reporting of potentially dangerous practices. These protections should be expanded to cover a wide range of potential whistleblowers, including employees, contractors, and external stakeholders who know of unsafe practices. Protection should include legal protection against retaliation, confidentiality, safe reporting channels, and the investigation of reports documenting unsafe practices.

Liability

The Roadmap emphasizes the need to "hold AI developers and deployers accountable if their products or actions cause harm to consumers." We agree that developers, deployers, and users should all be expected to behave responsibly in the creation, deployment, and use of AI systems, and emphasize that the imposition of liability on developers is particularly critical in order to encourage early design choices that prioritize the safety and wellbeing of the public.

The Roadmap also correctly points out that "the rapid evolution of technology and the varying degrees of autonomy in AI products present difficulties in assigning legal liability to AI companies and their users." Under

1 Throughout this document, where we recommend thresholds for the most advanced/dangerous GPAIS, we are generally referring to multi-metric quantitative thresholds set at roughly these levels. While the recent AI Executive Order ("Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence") presumes an AI model to be "dual-use" if is trained on 10^{26} FLOPs or more, we recommend a training compute threshold set at 10^{25} FLOPs to remain consistent with the EU AI Act's threshold for presuming systemic risk from GPAIS. Such a threshold would apply to fewer than 10 current systems, most of which have already demonstrated some capacity for hazardous capabilities.

current law, it is unclear who is responsible when an AI system causes harm, particularly given the complexity of the AI supply chain.

Joint and Several Liability

When an individual is harmed as a consequence of an AI system, there are several parties that could be responsible: the developer who trained the AI model, the provider who offers that model for use, the deployer who deploys the model as part of an AI system, or the user/consumer who employs the system for a given purpose. In addition, advanced GPAIS often serve as "foundation models," which are incorporated as one component of a more elaborate system, or which are fine-tuned by third-parties to select for particular characteristics. This presents the possibility for multiple parties to assume each of the aforementioned roles.

In such circumstances, joint and several liability is often appropriate. Joint and several liability provides that a person who has suffered harm can recover the full amount of damages from any of the joint and severally liable parties, i.e. those comprising the AI supply chain. The burden then rests on the defendant to recover portions of those damages from other parties based on their respective responsibilities for the harm. In other words, if a person is harmed by an AI system, that person would be able to sue any one of the developer, the provider, or the deployer of the system, and recover the full amount of damages, with these parties then determining their relative liability for that payment of damages independently of the injured party.

This absolves the person harmed of the burden of identifying the specific party responsible for the harm they suffered, which would be nearly impossible given the complexity of the supply chain, the opacity of the backend functions of these systems, and the likelihood that multiple parties may have contributed to the system causing harm. Instead, the defendant, who is more familiar with the parties involved in the system's lifecycle and the relative contributions of each to the offending function, would be charged with identifying the other responsible parties, and joining them as co-defendants in the case, as appropriate.

Strict Liability

Strict liability refers to a form of liability in which the exercise of due care is not sufficient to absolve a defendant of liability for harm caused by their action or product. While products are generally subject to a particular brand of strict liability, in most cases services rely on a negligence standard, which absolves the defendant of liability if due care was exercised in the conduct of the service. The lack of clarity as to whether advanced GPAIS should be classified as products or services draws into question whether this strict liability framework applies.

The inherent unpredictability of advanced GPAIS and inevitability of emergent unforeseen risks make strict liability appropriate. Many characteristics of advanced GPAIS render their training and provision akin to an "abnormally dangerous activity" under existing tort law, and abnormally dangerous activities are typically subject to strict liability. Existing law considers an activity abnormally dangerous and subject to strict liability if: (1) the activity creates a foreseeable and highly significant risk of physical harm even when reasonable care is exercised by all actors; and (2) the activity is not one of common usage.[2] A risk is considered highly significant if it is either unusually likely or unusually severe, or both. For instance, the operation of a nuclear power plant is considered to present a highly significant risk because while the likelihood of a harm-causing incident when reasonable care is exercised is low, the severity of harm should an incident occur would be extremely high.

2 Restatement (Third) of Torts: Liability for Physical Harm § 20 (Am. Law Inst. 1997).

A significant portion of leading AI experts have attested to a considerable risk of catastrophic harm from the most powerful AI systems, including executives from the major AI companies developing the most advanced systems.[3] The presence of these harms is thus evidently foreseeable and highly significant. Importantly, reasonable care is not sufficient to eliminate catastrophic risk from advanced GPAIS due to their inherent unpredictability and opacity, as demonstrated by the emergence of behaviors that were not anticipated by their developers in today's state-of-the-art systems.[4] As more capable advanced GPAIS are developed, this insufficiency of reasonable care will likely compound – an AGI system that exceeds human capacity across virtually all cognitive tasks, for instance, by definition would surpass the capacity of humans to exercise reasonable care in order to allay its risks.

Additionally, given the financial and hardware constraints on training such advanced models, only a handful of companies have the capacity to do so, suggesting that the practice also "is not one of common usage." In contrast, less capable systems are generally less likely to present emergent behaviors and inherently present a far lower risk of harm, particularly when reasonable care is exercised. Such systems are also less hardware intensive to train, and, while not necessarily "of common usage" at present, could qualify as such with continued proliferation.

Section 230

Section 230 of the Communications Decency Act of 1996 provides that, among other things, "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[5] This provision, along with the statute's protection of interactive computer services from liability for good faith moderation actions, has been broadly interpreted to protect online platforms from liability for the content they host, so long as the content was contributed by a party other than the platform itself.

The application of this statute to AI has yet to be tested in courts, and there is disagreement among legal scholars as to how Section 230 relates to generative AI outputs. On one hand, the outputs of generative AI systems are dependent on the input of another information content provider, i.e. the user providing the prompt. On the other hand, the outputs generated by the system are wholly unique, more akin to content provided by the platform itself. The prevailing view among academics is that generative AI products "operate on something like a spectrum between a retrieval search engine (more likely to be covered by Section 230) and a creative engine (less likely to be covered)."[6] A robust liability framework for AI must therefore ensure that this area of the law is clarified, either by explicitly superseding Section 230, or by amending Section 230 itself to provide this clarification. Shielding the developers and operators of advanced AI systems from liability for harms resulting from their products would provide little incentive for responsible design that minimizes risk, and, as we have seen with social media, could result in wildly misaligned incentives to the detriment of the American public.

3 In a survey of 2,778 researchers who had published research in top-tier AI venues, roughly half of respondents gave at least a 10% chance of advanced AI leading to outcomes as bad as extinction. K Grace, et al., "Thousands of AI Authors on the Future of AI," Jan. 2024, https://doi.org/10.48550/arXiv.2401.02843.

See also, e.g., S Mukherjee, "Top AI CEOs, experts raise 'risk of extinction' from AI," Reuters, May 30, 2023, https://www.reuters.com/technology/top-ai-ceos-experts-raise-risk-extinction-ai-2023-05-30/.

4 See, e.g., J Wei, et al., "Emergent Abilities of Large Language Models," Jun. 15, 2022 (last revised: Oct. 26, 2022), https://doi.org/10.48550/arXiv.2206.07682.

5 47 U.S.C. § 230(c)(1)

6 Henderson P, Hashimoto T, Lemley M. Journal of Free Speech Law. "Where's the Liability for Harmful AI Speech?" https://www.journaloffreespeechlaw.org/hendersonhashimotolemley.pdf#page=34

Recommendations

1. **The development of a GPAIS trained with greater than, e.g., 10^{25} FLOPs, or a system equivalent in capability[7], should be considered an abnormally dangerous activity and subject to strict liability** due to its inherent unpredictability and risk, even when reasonable care is exercised.
2. **Developers of advanced GPAIS that fall below this threshold, but still exceed a lower threshold (e.g. 10^{23} FLOPs) should be subject to a rebuttable presumption of negligence if the system causes harm** – given the complexity, opacity, and novelty of these systems, and the familiarity of their developers with the pre-release testing the systems underwent, developers of these systems are best positioned to bear the burden of proving reasonable care was taken.
3. **Domain-specific AI systems should be subject to the legal standards applicable in that domain.**
4. Where existing law does not explicitly indicate an alternative apportionment of liability, **AI systems should be subject to joint and several liability**, including for all advanced GPAIS.
5. **Congress should clarify that Section 230 of the Communications Decency Act does not shield AI providers from liability for harms resulting from their systems**, even if the output was generated in response to a prompt provided by a user. This could be accomplished by amending the definition of "information content provider" in Section 230(f)(3) to specify that the operator of a generative AI system shall be considered the "information content provider" for the outputs generated by that system.

AI and National Security

AI systems continue to display unexpected, emergent capabilities that present risks to the American public. Many of these risks, for example those from disinformation[8] and cyberattacks[9], were not discovered by evaluations conducted during the development life cycle (i.e. pre-training, training, and deployment), or were discovered but were deemed insufficiently severe or probable to justify delaying release. Moreover, AI companies have incentives to deploy AI systems quickly in order to establish and/or maintain market advantage, which may lead to substandard monitoring and mitigation of AI risks. When risks are discovered and harm is imminent or has occurred, it is vital for authorities to be informed as soon as possible to respond to the threat.

The Roadmap encourages committees to explore "whether there is a need for an AI-focused Information Sharing and Analysis Center to serve as an interface between commercial AI entities and the federal government to support monitoring of AI risks." We see such an interface as essential to preserving national security in light of the risks, both unexpected and reasonably foreseeable, presented by these systems.

The Roadmap also notes that "AI has the potential to increase the risk posed by bioweapons and is directly relevant to federal efforts to defend against CBRN threats." As state-of-the-art AI systems have become more advanced, they have increasingly demonstrated capabilities that could pose CBRN threats. For instance, in 2022, an AI system used for pharmaceutical research effectively identified 40,000 novel candidate chemical

7 See fn. 1.

8 Exclusive: GPT-4 readily spouts misinformation, study finds. Axios. https://www.axios.com/2023/03/21/gpt4-misinformation-newsguard-study

9 OpenAI's GPT-4 Is Capable of Autonomously Exploiting Zero-Day Vulnerabilities. Security Today. https://securitytoday.com/Articles/2024/04/23/OpenAIs-GPT4-Is-Capable-of-Autonomously-Exploiting-ZeroDay-Vulnerabilities.aspx

weapons in six hours[10]. While the current generation of models have not yet significantly increased the abilities of malicious actors to launch biological attacks, newer models are adept at providing the scientific knowledge, step-by-step experimental protocols, and guidance for troubleshooting experiments necessary to effectively develop biological weapons.[11] Additionally, currently models have been shown to significantly facilitate the identification and exploitation of cybervulnerabilities.[12] These capabilities are likely to scale over time.

Over the last two years, an additional threat has emerged at the convergence of biotechnology and AI, as ever-powerful AI models are 'bootstrapped' with increasingly sophisticated biological design tools , allowing for AI-assisted identification of virulence factors, in silico design of pathogens, and other capabilities that could significantly increase the capacity of malicious actors to cause harm.[13]

The US government should provide an infrastructure for monitoring these AI risks that puts the safety of the American public front and center, gives additional support to efforts by AI companies, and allows for rapid response to harms from AI systems. Several precedents for such monitoring already exist. For instance, CISA's Joint Cyber Defense Collaborative is a nimble network of cross-sector entities that are trusted to analyze and share cyber threats, the SEC requires publicly traded companies to disclose cybersecurity incidents within four business days, and the 2023 AI Executive Order requires companies to disclose 'the physical and cybersecurity protections taken to assure the integrity of that training process against sophisticated threats.'[14]

Recommendations

1. Congress should **establish an Information Sharing and Analysis Center (ISAC)** which will designate any model or system that meets a specified quantitative threshold[15] as a model or system of national security concern. Congress should **require developers building advanced AI systems to share documentation with the ISAC** about the decisions taken throughout the development and deployment life-cycle (e.g., models cards detailing decisions taken before, during, and after the training and release of a model).
2. The current draft of the 2025 National Defense Authorization Act (NDAA)[16] tasks the Chief Digital and Artificial Intelligence Officer with developing an implementation plan for a secure computing and data storage environment (an 'AIxBio sandbox') to facilitate the testing of AI models trained on biological data, as well as the testing of products generated by such models. **Congress should mandate that AI systems as or more powerful than those defined as models of national security concern (see above) or are otherwise deemed to pose CBRN threats be subjected to testing in this sandbox before deployment** to ensure that these systems do not pose severe risks to the American public. This type of faculty should

10 AI suggested 40,000 new possible chemical weapons in just six hours. The Verge. https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx

11 Can large language models democratize access to dual-use biotechnology? MIT. https://arxiv.org/ftp/arxiv/papers/2306/2306.03809.pdf

12 R Fang, et al., "Teams of LLM Agents can Exploit Zero-Day Vulnerabilities," Jun. 2, 2024, https://doi.org/10.48550/arXiv.2406.01637; T Claburn, "OpenAI's GPT-4 can exploit real vulnerabilities by reading security advisories," The Register, Apr. 17, 2024, https://www.theregister.com/2024/04/17/gpt4_can_exploit_real_vulnerabilities/ (accessed Jun. 13, 2024).

13 J O'Brien & C Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology," Health Secur., 2020 May/Jun; 18(3):219-227. doi: 10.1089/hs.2019.0122. https://pubmed.ncbi.nlm.nih.gov/32559154/.

14 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

SEC Adopts Rules on Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure by Public Companies. US Securities and Exchange Commission. https://www.sec.gov/news/press-release/2023-139

15 See fn. 1.

16 Draft text current as of May 25th, 2024.

follow design and protocols from the national security sector's Sensitive Compartmented Information Facility (SCIF) standards or the similar Data Cleanroom standards used in software litigation discovery.

3. To ensure that GPAIS are not capable of revealing hazardous information, **Congress should prohibit AI models from being trained on the most dangerous dual-use research of concern** (DURC). Congress should also **recommend appropriate restrictions for DURC data being used to train narrow AI systems** – such as ringfencing of the most hazardous biological information from use in training – that could pose significant risk of misuse, malicious use, or unintended harm. In both cases, these requirements should cover data that, if widely available, would pose a potential CBRN risk.
4. **The federal government should invest in core CBRN defense strategies that are agnostic to AI, while bearing in mind that AI increases the probability of these threats materializing.** Such investments should include next-generation personal protective equipment (PPE), novel medical countermeasures, ultraviolet-C technologies, and other recommendations from the National Security Commission for Emerging Biotechnology.[17]

Compute Security and Export Controls

High-end AI chips are responsible for much of the rapid acceleration in development of AI systems. As these chips are an integral component of AI development and rely on a fairly tight supply chain – i.e. the supply chain is concentrated in a small number of companies in a small number of countries[18] – chips are a promising avenue for regulating the proliferation of the highest-risk AI systems, especially among geopolitical adversaries and malicious non-state actors.

The Roadmap "encourages the relevant committees to ensure [the Bureau of Industry and Security] proactively manages [critical] technologies and to investigate whether there is a need for new authorities to address the unique and quickly burgeoning capabilities of AI, including the feasibility of options to implement on-chip security mechanisms for high-end AI chips." We appreciate the recognition by the AI Working Group of on-chip security as a useful approach toward mitigating AI risk. Congress must focus on both regulatory and technical aspects of this policy problem to mitigate the risk of AI development from malicious actors.

The Roadmap also asks committees to develop a framework for determining when, or if, export controls should be placed on advanced AI systems. We view hardware governance and export controls as complementary and mutually-reinforcing measures, wherein on-chip security mechanisms can serve to mitigate shortcomings of export controls as a means of reducing broad proliferation of potentially dangerous systems.

Export controls, especially those with an expansive purview, often suffer from serious gaps in enforcement. In response to export controls on high-end chips used for training AI systems, for instance, a growing informal economy around chip smuggling has already emerged, and is likely to grow as BIS restrictions on AI-related hardware and systems become more expansive.[19] Coupling export controls with on-chip governance mechanisms

17 Interim Report. National Security Commission on Emerging Biotechnology. https://www.biotech.senate.gov/press-releases/interim-report/ Also see AIxBio White Paper 4: Policy Options for AIxBio. National Security Commission on Emerging Biotechnology. https://www.biotech.senate.gov/press-releases/aixbio-white-paper-4-policy-options-for-aixbio/

18 Maintaining the AI Chip Competitive Advantage of the United States and its Allies. Center for Security and Emerging Technology. https://cset.georgetown.edu/wp-content/uploads/CSET-Maintaining-the-AI-Chip-Competitive-Advantage-of-the-United-States-and-its-Allies-20191206.pdf

19 Preventing AI Chip Smuggling to China. Center for a New American Security. https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china.

can help remedy this gap in enforcement by providing the ability to track and verify the location of chips, and to automatically or remotely disable their functionality based on their location when they are used or transferred in violation of export controls.

Export controls also generally target particular state actors rather than select applications, which may foreclose economic benefits and exacerbate geopolitical risks to United States interests relative to more targeted restrictions on trade. For example, broadly-applied export controls[20] targeted at the People's Republic of China (PRC) do not effectively distinguish between harmless use cases (e.g., chips used for video games or peaceful academic collaborations) and harmful use cases (e.g., chips used to train dangerous AI military systems) within the PRC. Expansive export controls have already led to severe criticism from the Chinese government,[21] and may be having the unintended effect of pushing China toward technological self-reliance.[22]

In contrast, relaxing restrictions on chip exports to demonstrably low-risk customers and for low-risk uses in countries otherwise subject to export controls could improve the economic competitiveness of US firms and strengthen trade relationships key to maintaining global stability. These benefits are integral to guaranteeing sustained US leadership on the technological frontier, and to maintaining the geopolitical posture of the US. The ability for on-chip governance mechanisms to more precisely identify the location of a given chip and to determine whether the chip is co-located with many other chips or used in a training cluster could facilitate more targeted export controls that maintain chip trade with strategic competitors for harmless uses, while limiting their application toward potentially risky endeavors.

New and innovative hardware governance solutions are entirely compatible with the current state of the art chips sold by leading manufacturers. All hardware relevant to AI development (i.e. H100s, A100s, TPUs, etc.) have some form of "trusted platform module (TPM)", a hardware device that generates random numbers, holds encryption keys, and interfaces with other hardware modules to ensure platform integrity and report security-relevant metrics.[23] Some new hardware (H100s in particular) has an additional "trusted execution environment (TEE)" or "secure enclave" capability, which prevents access to chosen sections of memory at the hardware level. TPMs and secure enclaves are already available and in use today, presently serving to prevent iPhones from being "jailbroken," or used when stolen, and to secure biometric and other highly sensitive information in modern phones and laptops. As discussed, they can also facilitate monitoring of AI development to identify the most concerning uses of compute and take appropriate action, including automatic or remote shutdown if the chips are used in ways or in locations that are not permitted by US export controls.

These innovations could be transformative for policies designed to monitor AI development, as TEEs and TPMs use cryptographic technology to guarantee confidentiality and privacy for all users across a variety of use and governance models.[24] Such guarantees are likely necessary for these chips to become the industry and international standard for use, and for willing adoption by strategic competitors. TEE and TPM security

20 BIS has limited information on distinguishing between use-cases, and is compelled to favor highly adversarial and broad controls to mitigate security risks from lack of enforcement.

21 China lashes out at latest U.S. export controls on chips. Associated Press. https://apnews.com/article/technology-business-china-global-trade-47eed4a9fa1c2f51027ed12cf929ff55

22 Examining US export controls against China. East Asia Forum. https://eastasiaforum.org/2024/03/16/examining-us-export-controls-against-china/

23 For more information on TPMs, see Safeguarding the Future of AI: The Imperative for Responsible Development. Trusted Computing Group. https://trustedcomputinggroup.org/safeguarding-the-future-of-ai-the-imperative-for-responsible-development/

24 Similar technology is also employed in Apple's Private Cloud Compute. See "Private Cloud Compute: A new frontier for AI privacy in the cloud," Apple Security Research Blog, Jun. 10, 2024, https://security.apple.com/blog/private-cloud-compute/.

capabilities can also be used to construct an "attested provenance" capability that gives cryptographic proof that a given set of AI model weights or model outputs results from a particular auditable combination of data, source code, training characteristics (including amount of compute employed), and input data. This provides a uniquely powerful tool in verifying and enforcing licensing standards.

Because state-of-the-art chips already possess the technical capability for this type of on-chip security, a technical solution to hardware governance would not impose serious costs on leading chip companies to modify the architecture of chips currently in inventory or in production. Additionally, it is possible to use these technical solutions for more centralized compute governance without creating back-channels that would harm the privacy of end-users of the chip supply chain – indeed these mechanisms can ensure privacy and limit communication of information to telemetry such as location and usage levels.

Recommendations

1. **Congress should support the passage of H.R.8315, the Enhancing National Frameworks for Overseas Restriction of Critical Exports (ENFORCE) Act**, which gives the Bureau of Industry and Security (BIS) the authority to control the export and re-export of covered AI systems, with amendments to ensure that the publication of AI models in a manner that is publicly accessible does not create a loophole to circumvent these controls, i.e., that open-weight systems meeting specified conditions qualify as exports under the Act.[25]
2. **Congress should require companies developing AI systems that meet specified thresholds26 to use AI chips with secure hardware.** This hardware should be privacy-preserving to allow for confidential computing but should also provide information on proof-of-location and the ability to switch chips off in emergency circumstances.[27] Such a technical solution would complement robust export controls by facilitating enforcement and more effectively targeting harmful applications in particular. This could be accomplished through direct legislation prohibiting the domestic training of advanced AI systems using chips without such technology, and by providing a statutory obligation for BIS to grant export licenses for high-end AI chips and dual-use AI models only if they are equipped with these on-chip security mechanisms and trained using such chips, respectively.
3. To avoid gaming, inaccuracy, or misrepresentation in the satisfaction of licensing requirements, **Congress should phase-in increasingly stringent evidentiary requirements for reporting of compute usage and auditing results.** The recently-established US AI Safety Institute within the National Institute of Standards and Technology should be tasked with developing a comprehensive standard for compute accounting to be used in threshold determinations. Additionally, self-attestation of compute usage and capability evaluations should be improved to cryptographically attested provenance when this becomes technically practical.

25 Open source systems developed in the United States have supercharged the development of AI systems in China, the UAE and elsewhere.

See, How dependent is China on US artificial intelligence technology? Reuters. https://www.reuters.com/technology/how-dependent-is-china-us-artificial-intelligence-technology-2024-05-09/;

Also see, China's Rush to Dominate A.I. Comes With a Twist: It Depends on U.S. Technology. New York Times. https://www.nytimes.com/2024/02/21/technology/china-united-states-artificial-intelligence.html

26 See fn. 1.

27 For an example of a technical project meeting these conditions, see the Future of Life Institute response to the Bureau of Industry and Security's Request for Comment RIN 0694–AI94 on implementation of additional export controls, which outlines an FLI project underway in collaboration with Mithril Security.

Autonomous Weapons Systems and Military Integration of AI

The Roadmap asks committees to take actions that prioritize the "development of secure and trustworthy algorithms for autonomy in DOD platforms" and ensure "the development and deployment of Combined Joint All-Domain Command and Control (CJADC2) and similar capabilities by DOD."

Following the 2021 CJADC2 Strategy[28], the Department of Defense (DOD) announced a new generation of capabilities for CJADC2 early this year, which intend to use AI to "connect data-centric information from all branches of service, partners, and allies, into a singular internet of military things." This built on similar efforts led by the Chief Digital and Artificial Intelligence Office (CDAO) and on the objectives of Task Force Lima to monitor, develop, evaluate, and recommend the responsible and secure implementation of generative AI capabilities across DOD.

While such innovations in the war-fighting enterprise present potential benefits – e.g., rapid integration of military intelligence, providing strategic decision advantage to commanders – there are significant pitfalls to rapid integration of AI systems, which have continued to be proven unreliable, opaque, and unpredictable. Bugs in AI systems used in such critical settings could severely hamper the national defense enterprise, and put American citizens and allies in danger, as a centralized system responsible for virtually all military functions creates a single point of failure and vulnerability. Integration of these systems may also lead to amplification of correlated biases in the decision-making of what would otherwise be independent AI systems used in military applications.

The Roadmap also "recognizes the DOD's transparency regarding its policy on fully autonomous lethal weapons systems [and encourages] relevant committees to assess whether aspects of the DOD's policy should be codified or if other measures, such as notifications concerning the development and deployment of such weapon systems, are necessary."

As the draft text of the 2025 National Defense Authorization Act (NDAA)[29] notes, the 'small unmanned aircraft systems (UAS) threat continues to evolve, with enemy drones becoming more capable and dangerous.' Autonomous weapons systems (AWS) are becoming increasingly cheap to produce and use, and swarms of such weapons pose a serious threat to the safety of citizens worldwide. When deployed en masse, swarms of autonomous weapons, which have demonstrated little progress in distinguishing between civilians and combatants in complex conflict environments, have the potential to cause mass casualties at the level of other kinds of WMDs. Their affordability also makes them a potentially potent tool for carrying out future genocides.

Overall, AWS have proven to be dangerously unpredictable and unreliable, demonstrating difficulty distinguishing between friend and foe. As these systems become more capable over time, they present a unique risk from loss of control or unintended escalation. Additionally, such systems are prone to cyber-vulnerabilities, and may be hacked by malicious actors and repurposed for malicious use.

Recommendations

1. **Congress should mandate that nuclear launch systems remain independent from CJADC2 capabilities.** The current air-gapped state of nuclear launch systems ensures that the critical decision to launch a nuclear weapon always remains within full human control. This situation also guards the nuclear command and

28 A CJADC2 Primer: Delivering on the Mission of "Sense, Make Sense, and Act". Sigma Defense. https://sigmadefense.com/wp-content/uploads/2023/09/CJADC2-White-Paper-Primer5.pdf

29 Draft text current as of May 25th, 2024.

control system against cyber-vulnerabilites which could otherwise present if the system was integrated with various other defense systems. Other systems may possess unique vulnerabilities from which nuclear launch systems are presently insulated, but to which they would be exposed were the functions integrated.

2. Building on the precedent set by the Air Force, **Congress should require DOD to establish boards comprised of AI ethics officers across all offices involved in the production, procurement, development, and deployment of military AI systems.**[30]
3. In light of the comments made by former Chief Digital and AI Officer Dr. Craig Martell that all AI systems integrated into defense operations must have 'five-digit accuracy' (99.999%),[31] **Congress should task the CDAO with establishing clear protocols to measure this accuracy and prohibit systems which fall below this level of accuracy from being used in defense systems.**
4. **Congress should codify DOD Directive 3000.09 in statute to ensure that it is firmly established, and amend it to raise the bar from requiring 'appropriate levels of human judgement' to requiring 'meaningful human control' when AI is incorporated in military contexts.** This is critical in ensuring that 'human-in-the-loop' is not used as a a rubber stamp, and in emphasizing the need for human control at each stage of deployment. In addition, Congress should require the CDAO to file a report which establishes concrete guidance for meaningful human control in practice, for both AWS and decision-support systems.
5. As the 2025 NDAA draft indicates, there is a need for development of counter-UAS (C-UAS) systems. Rather than ramping up development of unreliable and risky offensive AWS, **Congress should instead instruct DOD to invest in non-kinetic counter-AWS (C-AWS) development.** As AWS development accelerates and the risk of escalation heightens, the US should reassure allies that AWS is not the best countermeasure and instead push for advanced non-kinetic C-AWS technology.

Open-Source AI

Recently, "open-source AI" has been used to refer to AI models for which model weights, the numerical values that dictate how a model translates inputs into outputs, are widely available to the public. It should be noted that an AI system with widely available model weights alone does not fit the traditional criteria for open-source. The inconsistent use of this term has allowed many companies to benefit from the implication that models with varying degrees of openness might still fulfill the promises of open-source software (OSS), even when they do not adhere to the core principles of the open-source movement[32]. Contrary to the marketing claims of Big Tech companies deploying "open-source" AI models, Widder et al. (2023) argue that while maximally "open" AI can indeed provide transparency, reusability, and extensibility, allowing third parties to deploy and build upon powerful AI models, it does not guarantee democratic access, meaningful competition, or sufficient oversight and scrutiny in the AI field.

Advanced AI models with widely available model weights pose particularly significant risks to society due to their unique characteristics, potential for misuse, and the difficulty of evaluating and controlling their capabilities. In the case of CBRN risks, as of early 2024, evidence suggests that the current generation of closed AI systems function

30 "Air Force names Joe Chapa as chief responsible AI ethics officer". FedScoop. https://fedscoop.com/joe-chapa-air-force-chief-responsible-ai-ethics-officer/

31 "US DoD AI chief on LLMs: 'I need hackers to tell us how this stuff breaks." Venture Beat. https://venturebeat.com/ai/us-dod-ai-chief-on-llms-i-need-hackers-to-tell-us-how-this-stuff-breaks/

32 Widder, David Gray and West, Sarah and Whittaker, Meredith, Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI (August 17, 2023). Available at SSRN: https://ssrn.com/abstract=4543807

as instruments comparable to internet search engines in facilitating the procurement of information that could lead to harm.[33] However, these experiments were carried out using proprietary models with fine-tuned safeguards. The release of model weights allows for trivial removal of any safeguards that might be added to mitigate these risks and lowers the barrier to entry for adapting systems toward more dangerous capabilities through fine-tuning.[34, 35, 36]

As AI models become more advanced, their reasoning, planning, and persuasion capabilities are expected to continue to grow, which will in turn increase the potential for misuse by malicious actors and loss of control over the systems by careless operators. Relevant legislation should account for the difficulty in accurately predicting which models will possess capabilities strong enough to pose significant risks with and without the open release of their model weights.[37] Unanticipated vulnerabilities and dangerous capabilities can be particularly insidious in the latter case, as once model weights are released, such models cannot be effectively retracted in order patch issues, and the unpatched versions remain indefinitely available for use.

"Open AI systems" have already demonstrated the potential to facilitate harmful behavior, particularly by way of cyberattacks, disinformation, and the proliferation of child sexual abuse material (CSAM).[38, 39] The UK National Cyber Security Centre found that AI systems are expected to significantly increase the volume and impact of cyber attacks by 2025, with varying degrees of influence on different types of cyber threats.[40] While the near-term threat primarily involves the enhancement of existing tactics, techniques, and procedures, AI is already being used by both state and non-state actors to improve reconnaissance and social engineering. More advanced AI applications in cyber operations would likely to be limited to well-resourced actors with access to quality training data, immense computational resources, and expertise, but open release of model weights by these well-resourced actors could provide the same capacity to a wider range of threat actors, including cybercriminals and state-sponsored groups.

The Roadmap asks committees to "investigate the policy implications of different product release choices for AI systems, particularly to understand the differences between closed versus fully open-source models (including the full spectrum of product release choices between those two ends of the spectrum)." We appreciate the Roadmap's implication that "open-source model" product releases present additional questions in understanding the risks posed by AI systems, and recommend the following measures to mitigate the unique risks posed by the release of model weights.

Recommendations

1. **Congress should require that AI systems with open model weights undergo thorough testing and evaluation in secure environments appropriate to their level of risk.** The government should conduct these assessments directly or delegate them to a group of government-approved independent auditors.

33 Mouton, Christopher A., Caleb Lucas, and Ella Guest, The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. Santa Monica, CA: RAND Corporation, 2024. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

34 Lermen, Simon, Charlie Rogers-Smith, and Jeffrey Ladish. "LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B." arXiv, Palisade Research, 2023, https://arxiv.org/abs/2310.20624.

35 Gade, Pranav, et al. "BadLlama: Cheaply Removing Safety Fine-Tuning from Llama 2-Chat 13B." arXiv, Conjecture and Palisade Research, 2023, https://arxiv.org/abs/2311.00117.

36 Yang, Xianjun, et al. "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models." arXiv, 2023, https://arxiv.org/abs/2310.02949.

37 M Anderljung, et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety." Nov. 7, 2023. pp.35-36. https://arxiv.org/pdf/2307.03718 [accessed June 7, 2024].

38 CrowdStrike. 2024 Global Threat Report. CrowdStrike, 2023, https://www.crowdstrike.com/global-threat-report/.

39 Thiel, David, Melissa Stroebel, and Rebecca Portnoff. "Generative ML and CSAM: Implications and Mitigations." Stanford Cyber Policy Center, Stanford University, 24 June 2023, https://cyber.fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations.

40 The near-term impact of AI on the cyber threat. (n.d.). https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

When assessing these models, auditors must assume that a) built-in safety measures or restrictions could be removed or bypassed once the model is released, and b) the model could be fine-tuned or combined with other resources, potentially leading to the development of entirely new and unanticipated capabilities. Insufficient safeguards to protect against dangerous capabilities or dangerous unpredictable behavior should justify the authority to suspend the release of model weights, and potentially the system itself, until such shortcomings are resolved. In cases where full access to a model's weights is needed to evaluate reliably audit the capabilities of a system, assessment should be conducted in Sensitive Compartmented Information Facilities (SCIFs) to ensure appropriate security measures.[41],[42]

2. **Developers should be legally responsible for performing all reasonable measures to prevent their models from being retrained to substantially enable illegal activities, and for any harms resulting from their failure to do so.** When model weights are made widely available, it becomes intractable for developers to retract, monitor, or patch the system. Presently, there is no reliable method of comprehensively identifying all of the capabilities of an AI system. Latent capabilities, problematic use-cases, and vulnerabilities are often identified far into the deployment life-cycle of a system or through additional fine-tuning. Despite the difficulties in identifying the full range of capabilities, developers should be held liable if their model was used to substantially enable illegal activities.
3. **To mitigate the concentration of power of AI while ensuring AI safety and security, initiatives like the National Artificial Intelligence Research Resource (NAIRR) should be pursued to create "public options" for AI.** As previously discussed, the impacts of open-source AI on the concentration of power and on mitigating market consolidation are often overstated. This does not discount the importance of preventing the concentration of power, both within the technology market and for society at large, that is likely to result from the high barrier to entry for training the most advanced AI systems. One potential solution is for the U.S. to further invest in "public options" for AI. Initiatives like the National Artificial Intelligence Research Resource could help develop and maintain publicly-funded AI models, services, and infrastructure. This approach would ensure that access to advanced AI is not solely controlled by corporate or proprietary interests, allowing researchers, entrepreneurs, and the general public to benefit from the technology while prioritizing safety, security, and oversight.

Supporting US AI Innovation

The Roadmap has the goal of "reaching as soon as possible the spending level proposed by the National Security Commission on Artificial Intelligence (NSCAI) in their final report: at least $32 billion per year for (non-defense) AI innovation." We appreciate the support for non-military innovation of AI, and emphasize that AI innovation should not be limited to advancing the capabilities or raw power of AI systems. Rather, innovation should prioritize specific functions that maximize public benefit and tend to be under-incentivised in industry, and should include extensive research into improving the safety and security of AI systems. This means enhancing explainability of outputs, tools for evaluation of risk, and mechanisms for ensuring predictability and maintenance of control over system behavior.

41 S Casper, et al., "Black-Box Access is Insufficient for Rigorous AI Audits," Jan. 25, 2024 (last revised: May 29, 2024), https://doi.org/10.1145/3630106.3659037.

42 Editor, C. C. (n.d.). Sensitive Compartmented Information Facility (SCIF) – glossary: CSRC. CSRC Content Editor. https://csrc.nist.gov/glossary/term/sensitive_compartmented_information_facility

To this end, the Roadmap also expresses the need for funding efforts to enhance AI safety and reliability through initiatives to support AI testing and evaluation infrastructure and the US AI Safety Institute, as well as increased resources for BIS to ensure effective monitoring and compliance with export control regulations. The Roadmap also emphasizes the importance of R&D and interagency coordination focused on the intersection of AI and critical infrastructure. We commend the comprehensive approach to R&D efforts across multiple agencies, as it recognizes the critical role that each of these entities plays in ensuring the safe and responsible development of AI technologies. In particular, we see the intersection of AI and critical infrastructure as a major vector of potential AI risk if due care is not paid to ensuring reliability and security of systems integrated in critical infrastructure, and in strengthening resilience against possible AI-assisted cyberthreats.

Research focused on the safe development, evaluation, and deployment of AI is vastly under-resourced when compared to research focused on the general development of AI. AI startups received almost $50 billion in funding in 2023.[43] According to the 2024 Stanford Index Report, industry produced 51 notable machine learning models, academia contributed 15, and the government contributed 2.[44] While the amount of resources that private companies allocate to safety research is unclear – there can be some overlap between safety and capabilities research – it is significantly less than investment in capabilities. Recently, the members of teams working on AI safety at OpenAI have resigned citing concerns about the company's approach to AI safety research.[45] This underscores the need for funding focused on the safe development, evaluation, and deployment of AI.

Recommendations

1. **R&D funding to BIS should include allocation to the development of on-chip hardware governance solutions, and the implementation of those solutions.** To best complement the role of BIS in implementing export controls on advanced chips and potentially on AI models, this funding should include R&D supporting the further development of privacy-preserving monitoring such as proof-of-location and the ability to switch chips off in circumstances where there is a significant safety or regulatory violation.[46] After appropriate on-chip governance solutions are identified, funding should also be directed towards enabling the implementation of those solutions in relevant export control legislation.
2. **The expansion of NAIRR programs should include funding directed toward the development of secure testing and usage infrastructure for academics, researchers, and members of civil society.** We support efforts by the NAIRR pilot program to improve public access to research infrastructure. As AI systems become increasingly capable, levels of access to AI tools and resources should be dynamic relative to their level of risk. Accordingly, it may be beneficial for those receiving any government funding for their work on powerful models (including private sector) to provide structured access to their systems via the NAIRR, subject to specific limitations on use and security measures, including clearance and SCIFs, where necessary, to allow for third parties to probe these systems and develop the tools necessary to make them safer.

43 "Rounds Raised by Startups Using AI In 2023," Crunchbase, https://www.crunchbase.com/lists/rounds-raised-by-startups-using-ai-in/47342cf0-88af-4541-b01e-5c82f6f2d0bb/funding_rounds.

44 N Maslej, et al., "The AI Index 2024 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, Apr. 2024, https://aiindex.stanford.edu/report/.

45 Roose, K. (2024, June 4). OpenAI insiders warn of a "reckless" race for dominance. The New York Times. https://www.nytimes.com/2024/06/04/technology/openai-culture-whistleblowers.html

46 For an example of a technical project meeting these conditions, see the Future of Life Institute response to the Bureau of Industry and Security's Request for Comment (RIN 0694–AI94), which outlines an FLI project underway in collaboration with Mithril Security. Additional detail on the implementation of compute governance solutions can be found in the "Compute Security and Export Controls" section of this document.

3. **R&D on interagency coordination focused on the intersection of AI and critical infrastructure should include allocation to safety and security research.** The ultimate goal of this research should be to establish stringent baseline standards for the safe and secure integration of AI into critical infrastructure. These standards should address key aspects such as transparency, predictability, and robustness of AI systems, ensuring that they can be effectively integrated without introducing additional vulnerabilities. Funding should also acknowledge the lower barrier to entry for malicious actors to conduct cyberattacks as publicly-accessible AI becomes more advanced and widespread, and seek improved mechanisms to strengthen cybersecurity accordingly.

Combating Deepfakes

The Roadmap encourages the relevant committees to consider legislation "to protect children from potential AI-powered harms online by ensuring companies take reasonable steps to consider such risks in product design and operation." We appreciate the Roadmap's recognition that product design, and by extension product developers, play a key role in mitigating AI-powered harms. The Roadmap also encourages the consideration of legislation "that protects against unauthorized use of one's name, image, likeness, and voice, consistent with First Amendment principles, as it relates to AI,""legislation to address online child sexual abuse material (CSAM), including ensuring existing protections specifically cover AI-generated CSAM," and "legislation to address similar issues with non-consensual distribution of intimate images and other harmful deepfakes."

Deepfakes, which are pictures, videos, and audio that depict a person without their consent, usually for the purpose of harming that person or misleading those who are exposed to the material, lie at the intersection of these objectives. There are many ways in which deepfakes systematically undermine individual autonomy, perpetuate fraud, and threaten our democracy. For example, 96% of deepfakes are sexual material[47] and fraud committed using deepfakes rose 3,000% globally in 2023 alone.[48] Deepfakes have also begun interfering with democratic processes by spreading false information and manipulating public opinion[49] through convincing fake media, which can and have influenced electoral outcomes[50]. The Roadmap encourages committees to "review whether other potential uses for AI should be either extremely limited or banned." We believe deepfakes fall into that category.

Deepfakes are the result of a multilayered supply chain, which begins with model developers, who design the underlying algorithms and models. Cloud compute providers such as Amazon Web Services (AWS), Google Cloud Platform, and Microsoft Azure form the next link in the chain by offering the necessary computational resources for running and in some cases training deepfake models. These platforms provide the infrastructure and scalability required to process large datasets and generate synthetic media efficiently. Following them are

47 H Ajder et al. "The State of Deepfakes." Sept. 2019. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

48 Onfido. "Identity Fraud Report 2024." 2024. https://onfido.com/landing/identity-fraud-report/

49 G De Vynck. "OpenAI finds Russian and Chinese groups used its tech for propaganda campaigns." May. 30, 2024. https://www.washingtonpost.com/technology/2024/05/30/openai-disinfo-influence-operations-china-russia/

50 M Meaker. "Slovakia's election deepfakes show AI Is a danger to democracy." Oct. 3, 2023. https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/

model providers, such as Deepnude[51], Deepgram[52], and Hoodem[53], which offer access to pre-trained deepfake models or user-friendly software tools, enabling even those with limited technical expertise to produce deepfakes.

The end users of deepfake technology are typically individuals or groups with malicious intent, utilizing these tools to spread misinformation, manipulate public opinion, blackmail individuals, or engage in other illicit activities. Once created, these deepfakes are distributed through various online platforms, including social media sites such as Facebook, Twitter, and YouTube, as well as messaging apps like WhatsApp and Telegram. The proliferation of deepfakes on these platforms can be rapid and extensive, making it nearly impossible to remove the synthetic media once published. Accordingly, it is critical to prevent the production of deepfakes before their publication and distribution can occur.

As the creators and distributors of the powerful tools that enable the mass production of this harmful content, model developers and providers hold the most control and responsibility in the deepfake supply chain. Developers have the capability to stop the misuse of these technologies at the source by restricting access, disabling harmful functionalities, and simply refusing to train models for harmful and illegal purposes such as the generation of non-consensual intimate images. There are far fewer model developers than providers, making this link in the supply chain particularly effective for operationalizing accountability mechanisms. While compute providers also play a role by supplying the necessary resources for AI systems to function, their ability to monitor and control the specific use cases of these resources is more limited. In order to effectively stem off the risks and harms which deepfakes engender, legislative solutions must address the issue as a whole, rather than only in particular use cases, in order to reflect the broad and multifaceted threats that extend beyond any single application. A comprehensive legal framework would ensure that all potential abuses are addressed, creating a robust defense against the diverse and evolving nature of deepfake technology.

Recommendations

1. Congress should set up accountability mechanisms that reflect the spread of responsibility and control across the deepfake supply chain. Specifically, **model developers and providers should be subject to civil and/or criminal liability for harms resulting from deepfakes generated by their systems.** Similar approaches have been taken in existing Congressional proposals such as the NO AI FRAUD Act (H.R. 6943), which would create a private right of action against companies providing a "personalized cloning service". When a model is being used to quickly and cheaply create an onslaught of deepfakes, merely holding each end-user accountable would be infeasible and would nonetheless be insufficient to prevent the avalanche of harmful deepfakes flooding the internet.
2. **Users accessing models to produce and share deepfakes should be subject to civil and/or criminal liability.** This approach is already reflected in several bills proposed within Congress such as the NO AI FRAUD Act (H.R. 6943), NO FAKES Act, DEFIANCE Act (S.3696), and the Preventing Deepfakes of Intimate Images Act (H.R. 3106).
3. **Congress should place a responsibility on compute providers to revoke access to their services when they have knowledge that their services are being used to create harmful deepfakes, or to host models**

51 S Cole. "This horrifying app undresses a photo of any Woman with a single click." Jun. 26, 2019. https://www.vice.com/en/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman

52 Deepgram. "Build voice into your apps." https://deepgram.com/

53 Hoodem. "Create any deepfake with no limitation." https://hoodem.com/

that facilitate the creation of harmful deepfakes. This will ensure that compute providers are not complicit in the mass production of deepfakes.

4. **Congress should support the passage of proposed bills like the NO FAKES Act, with some modifications to clarify the liability of model developers.**[54] Many recently introduced bills contain elements which would be effective in combating deepfakes, although it is crucial that they are strengthened to adequately address the multilayered nature of the deepfakes supply chain.

Provenance and Watermarking

Watermarking aims to embed a statistical signal into AI-generated content, making it identifiable as such. Ideally, this would allow society to differentiate between AI-generated and non-AI content. However, watermarking has significant drawbacks. First, deepfakes such as non-consensual intimate images and CSAM are still considered harmful even when marked as AI-generated.[55] Websites hosting AI-generated sexual images often disclose their origin, yet the content continues to cause distress to those depicted. Second, recent research has shown that robust watermarking is infeasible, as determined adversaries can easily remove these markers.[56] As such, it is not sufficient to rely on watermarking alone as the solution to preventing the proliferation of deepfakes, nor for conclusively distinguishing real from synthetic content.

Nonetheless, certain types of watermarks and/or provenance data can be beneficial to combating the deepfake problem. "Model-of-origin" watermarking provisions, which would require generative AI models to include information on which model was used to create the output and the model's developer and/or provider, can be included in the metadata of the output and can greatly enhance both legal and public accountability for developers of models used to create harmful content. Indicating the model of origin of outputs would also enable the identification of models that are disproportionately vulnerable to untoward use.

Consistent with this approach, the Roadmap encourages committees to "review forthcoming reports from the executive branch related to establishing provenance of digital content, for both synthetic and non-synthetic content." It also recommends considering "developing legislation that incentivizes providers of software products using generative AI and hardware products such as cameras and microphones to provide content provenance information and to consider the need for legislation that requires or incentivizes online platforms to maintain access to that content provenance information."

While forthcoming legislation should indeed require providers of AI models to include content provenance information embedded in or presented along with the outputted content, however, developers should also bear this responsibility. Unlike model providers, developers can embed provenance information directly into the models during the development phase, ensuring that it is an integral part of the AI-generated content from the outset.

54 See Future of Life Institute's 'Recommended Amendments to Legislative Proposals on Deepfakes' report: https://futureoflife.org/document/recommended-amendments-to-legislative-proposals-on-deepfakes/

55 M B Kugler, C Pace. "Deepfake privacy: Attitudes and regulation." Feb. 8, 2021. https://www.scholars.northwestern.edu/en/publications/deepfake-privacy-attitudes-and-regulation

56 H Zhang, B Elderman, & B Barak. "Watermarking in the sand." Nov. 9, 2023. Kempner Institute, Harvard University. https://www.harvard.edu/kempner-institute/2023/11/09/watermarking-in-the-sand/

Recommendations

1. **Both model developers and providers should be required to integrate provenance tracking capabilities into their systems.** While voluntary commitments have been made by certain developers, provenance watermarking is most trustworthy when it is widespread, and this is not currently the industry norm. As the National Institute of Standards and Technology report on Reducing Risks Posed by Synthetic Content (NIST AI 100-4) outlines, several watermarking and labeling techniques have become prominent, meaning that there are established standards that can be viably adopted by both developers and providers.
2. **Model developers and providers should be expected to make content provenance information as difficult to bypass or remove as possible, taking into account the current state of science.** It is unlikely that most users creating deepfakes have the technical competency to remove watermarks and/or metadata, but model-of-origin provisions are nonetheless most effective if they are inseparable from the content. While studies have shown that malicious actors can bypass current deepfake labeling and watermarking techniques, stakeholders should ensure, to the greatest extent possible, that such bypasses are minimized. The absence of requirements that model-of-origin information be as difficult to remove as possible may unintentionally incentivize developers and deployers to employ watermarks that are easier to remove in an effort to minimize accountability.
3. **Congress should support the passage of the AI Labeling Act, which mandates clear and permanent notices on AI-generated content, identifying the content as AI-produced and specifying the tool used along with the creation date.** This transparency helps hold developers accountable for harmful deepfakes, potentially deterring irresponsible AI system design.
4. **Congress should support amendments to various bills originating in the House, including the AI Disclosure Act and the DEEPFAKES Accountability Act, such that they clearly include model-of-origin watermarking provisions.**[57]

57 See Future of Life Institute's 'Recommended Amendments to Legislative Proposals on Deepfakes' report: https://futureoflife.org/document/recommended-amendments-to-legislative-proposals-on-deepfakes/

Conclusion

We thank the Senate AI Working Group for its continued dedication to the pressing issue of AI governance. AI as a technology is complex, but the Roadmap demonstrates a remarkable grasp of the major issues it raises for the continued flourishing of the American people. The next several months will be critical for maintaining global leadership in responsible AI innovation, and the urgent adoption of binding regulation is essential to creating the right incentives for continued success. Time and time again, Congress has risen to meet the challenge of regulating complex technology, from airplanes to pharmaceuticals, and we are confident that the same can be done for AI.