A Path for Science- and Evidence-based AI Policy

Rishi Bommasani[*1], Sanjeev Arora[3], Yejin Choi[4], Daniel E. Ho[1], Dan Jurafsky[1], Sanmi Koyejo[1], Hima Lakkaraju[5], Fei-Fei Li[1], Arvind Narayanan[3], Alondra Nelson[6], Emma Pierson[7], Joelle Pineau[8], Gaël Varoquaux[9], Suresh Venkatasubramanian[10], Ion Stoica[2], Percy Liang[1], Dawn Song[*2]

[1]Stanford University [2]UC Berkeley [3]Princeton University [4]University of Washington [5]Harvard University [6]Institute for Advanced Study [7]Cornell University [8]McGill University [9]INRIA [10]Brown University

---

Overview: AI is a powerful technology that carries both benefits and risks. We wish to promote innovation to ensure its potential benefits are responsibly realized and widely shared, while simultaneously ensuring that current and potential societal risks are mitigated. To address the growing societal impact of AI, many jurisdictions are pursuing policymaking. The AI research and policy community lacks consensus on the evidence base relevant for effective policymaking, as has been seen with the debates over California's Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (California's SB-1047). Points of contention include disagreement about what risks should be prioritized, if or when they will materialize, and who should be responsible for addressing these risks.

In light of this, we firmly believe AI policy should be informed by scientific understanding of AI risks and how to successfully mitigate them. Therefore, if policymakers pursue highly committal policy, the evidence of the associated AI risks should meet a high evidentiary standard. Advancing significant legislation without clear understanding of risks it intends to address may lead to more negative consequences than positive outcomes.

We support evidence-based policy and recognize current scientific understanding is quite limited. Therefore, we recommend the following priorities to advance scientific understanding and science- and evidence-based AI policy:

- We need to better understand AI risks.
- We need to increase transparency on AI design and development.
- We need to develop techniques and tools to actively monitor post-deployment AI harms and risks.
- We need to develop mitigation and defense mechanisms for identified AI risks.
- We need to build trust and reduce fragmentation in the AI community.

We describe each of these priorities in more detail below. We believe by following these steps we can pave a more productive path toward robust and responsible AI, anchored in the best scientific practice and AI policymaking that is evidence based.

First, we need to better understand AI risks. As AI models rapidly advance in capability, our level of investment dedicated to research for understanding risks should also increase. However, our understanding of how these models function and their possible negative impacts on society remains very

limited. Because this technology has the potential to be far more powerful than existing technologies and carries wide-reaching implications, it is crucial to have a comprehensive understanding of AI risks. Such a comprehensive understanding of AI risks is the necessary foundation for effective policy. In particular, we recommend extensive and comprehensive study of a broad spectrum of different risks, including discrimination, scams, misinformation, non-consensual intimate imagery, child sexual abuse material, cybersecurity risks, environmental risks, biosecurity risks, and extreme risks. To build understanding, we recommend such studies apply a marginal risk framework, assessing the additional societal risks posed by AI models compared to existing technologies like internet search engines. This approach will identify AI's unique risks and negative impacts on society. Given our current limited understanding of risks, we recommend policymakers invest resources across academia, civil society, government and industry to foster research on AI risk analysis. Ultimately, these risk analyses will help inform decisions on important policy questions such as whether to release a model, how to release it, and what uses of AI should be permitted.

Second, we need to increase transparency. Analyzing risks and developing policy is challenging without adequate information. Transparency requirements for advanced models would facilitate risk analysis of capable models with potentially significant impacts. A number of important questions related to transparency require further study. One key question is the criteria used in policymaking to determine which entities and models are in scope. Current policies such as the US Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence and the European Union's AI Act set thresholds based on compute used to train an AI model, but there is currently no solid evidence behind these numbers. More rigorous studies should be required to establish criteria that strike the right balance between innovation and risk mitigation. Furthermore, we need to study the level of information disclosure that best balances usefulness and potential overhead for model developers. Key questions for increasing transparency include: what information should be shared with different parties (e.g. the public, trusted third parties, the government); what information developers should report about models including size, a summary of training data and methods, capabilities (such as test results on certain benchmarks and red teaming practices); and incidents such as model theft, unauthorized access, and the inadvertent release of model weights. In line with this approach, the establishment of a registry for capable models might improve transparency.

Third, we need to develop an early warning detection mechanism. While current model capabilities may not pose devastating consequences, given the rapid pace of technological development, significantly more powerful and risky AI models are likely to emerge in the near future. It's crucial to draw on research to establish early warning detection mechanisms as soon as possible, providing society with more time to implement stronger mitigation and response measures, and hopefully preventing devastating consequences and the crossing of predefined red lines. The mechanisms should encompass both in-lab testing and real-world monitoring. In a lab setting, we need to detect and evaluate the dangerous capabilities of models through rigorous evaluation and red teaming. This involves testing AI models with adversarial scenarios to uncover potential vulnerabilities or unintended behaviors, and assessing how the models could lead to significant marginal risks in areas such as cybersecurity and biochemical weapons. Such evaluation and red teaming analyses can help identify risks before models are deployed in the real world. After the models are launched, the mechanism should enable adverse event reporting, documenting how AI has been used for misuse and what consequences it has caused in the real world. To this end, continuous real-world monitoring of how misuse of AI may have caused harm in each application domain

such as biotechnology and cybersecurity is crucial. Moreover, we need to determine to whom these early warnings should be reported and design a responsible report protocol. This will help ensure that identified risks are communicated effectively among the relevant authorities or stakeholders.

Fourth, we need to develop technical mitigation and defense mechanisms. Investing in research to create these solutions is crucial for effectively addressing a wide range of AI risks. First, instead of just relying on today's alignment approaches, it is important to explore and develop new approaches for building safe AI with the potential for greater safety assurance. This is a complex challenge, which requires a multifaceted approach. Second, in addition to the long-term research that might be required to develop safer AI models, it is also important to develop defensive approaches or immune systems in society to reduce the potential negative impacts from misuse of AI technology. For example, additional defensive systems can involve improving the security posture and defenses of computer systems against security risks caused by AI misuse.

Fifth, we need to build trust and unite the community by reducing fragmentation. Currently, the AI community is heavily fragmented with a variety of views on approaches to risk and policy: the fragmentation poses a challenge to scientific consensus that would support evidence-based AI policymaking. One extreme position calls for strong regulation to mitigate extreme risks, whereas the other extreme calls for no regulation to avoid inhibiting innovation. We support a third approach, an evidence-based approach to AI policy, which reduces fragmentation towards finding the best solutions for fostering innovation while mitigating risks. To achieve this, we recommend the continued development of collaborative research initiatives that bring together diverse perspectives. Creating platforms for respectful scientific debate and shared problem-solving in a trusted setting can help bridge the gap between different viewpoints. Additionally, establishing interdisciplinary research groups that include AI researchers, social scientists, legal scholars, domain experts, policymakers, and industry representatives can promote a more inclusive and comprehensive approach to AI development, AI safety, and AI policy. We hope this approach can also lead to greater international cooperation.

Call to action.

We call upon the AI research and policy communities to proactively work together to advance science and develop evidence-based policy.

We envision that these communities could produce a forward-looking design, or *blueprint*, for AI policy that maps different conditions that may arise in society (e.g. specific model capabilities, specific demonstrated harms) to candidate policy responses. By working together on this practical, actionable blueprint for the future, we can work towards scientific consensus, even when different stakeholders significantly disagree on how capabilities will evolve or how risks should be addressed. Such a blueprint would complement current guidance on how AI can be developed and deployed responsibly such as the National Institute of Standards and Technology Risk Management Framework use-case profiles.

A blueprint for the appropriate conditional response under different societal conditions would organize the discourse on AI policymaking around a concrete artifact. In turn, even if consensus cannot be reached on the likelihood of particular outcomes (e.g. how much can AI increase the likelihood of large-scale cyber attacks or disinformation campaigns), progress could be made towards determining the appropriate course of action.

We should organize a series of convenings as a forum for sustained multi-stakeholder dialogue that reflects many different positions, disciplines, and institutions. Through these convenings, we will develop milestones toward progress to science-based AI policy recommendations:

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

By taking these steps, we will build a strong evidentiary foundation, along with broad engagement and thoughtful deliberation, for producing better policy.

September 24, 2024