Policy supplement:
Managing AI Risks in an Era of Rapid Progress

This supplement highlights and expands on key policy recommendations from the paper. It was prepared by a subset of the authors of the original paper, and is not a complete summary.

Executive summary

Experts call attention to growing dangers from the malicious use of AI and premature reliance on unreliable AI systems.

In a recent short paper, world-leading AI scientists and governance experts from the US, China, EU, UK, and other countries have highlighted that rapid AI progress will pose societal-scale risks.

Along with their benefits, today's AI systems already contribute to a wide array of harms, from eroding social trust to enabling criminals and terrorists. And over the coming years, the best-funded AI companies plan to pour billions of dollars into building far more capable AI systems. Meanwhile, other institutions may face pressure to adopt flawed AI systems without understanding their downsides.

Due to their greater capabilities and their potential deployment in many industries, future AI systems will pose many risks to society. These risks include rapid job displacement, automated misinformation, and enabling large-scale cyber and biological threats. Experts are also concerned that labs could lose control over frontier systems as these systems become increasingly good at coding, planning, and persuasion.

Proposed policy measures for labs and governments:

1. **Both industry and government should invest one-third of their AI R&D resources into research on safe, ethical AI.**

2. **Industry and governments should set standards for assessing and mitigating the risks of large AI models.**

3. **Governments should establish oversight, monitoring, and liability** for the AI industry.

4. **Governments should take further preparatory measures against emerging risks** from frontier AI systems.

Policy Recommendations

1. Industry labs and government funders should invest in safe, ethical AI.

Industry and government should allocate at least one-third of their AI R&D resources[1] to ensure the safety and ethical use of AI systems. Relevant research areas include:

1. **Oversight and honesty:** More capable AI systems are better able to exploit weaknesses in oversight and testing—for example, by producing false but compelling output.

2. **Robustness:** AI systems behave unpredictably in new situations (under distribution shift or adversarial inputs).

3. **Interpretability:** AI decision-making is opaque. So far, we can only test large models via trial and error. We need to learn to understand their inner workings.

4. **Risk evaluations:** Frontier AI systems often develop capabilities their creators don't expect, which may be discovered late in their training or even well after deployment. Better evaluation is needed to detect or predict hazardous capabilities as early as possible.

5. **Addressing emerging challenges:** More capable future AI systems may exhibit failure modes we have so far seen only in theoretical models. AI systems might, for example, learn to feign obedience or exploit weaknesses in our safety objectives and shutdown mechanisms to advance a particular goal.

This list is not exhaustive. Additional relevant R&D areas are described in Hendrycks et al. Further, Hendrycks and Mazeika define one class of safety-adjacent activities which should *not* count as ensuring safety: activities that do not improve the *safety-capabilities balance* because they accelerate general AI capabilities as much or more than they improve safety metrics.

2. Industry and governments should establish policies and set standards for assessing and mitigating the risks of large AI models.

1. **Frontier AI developers should promptly commit to detailed and independently scrutinized scaling policies**. These policies should describe specific safety measures that these companies will take if specific dangerous capabilities are found in their AI systems.

2. **Governments should set national and international safety standards for AI training and deployment.**

   (a) **Standards should identify practices to address a variety of risk vectors,** including misuse of publicly available models, theft of models in development, accidents and incidents from unintended behavior, and societal effects of the widespread use of unreliable models.

[1] "Resources" here includes both funding and talent.

    (b) **Standards should be responsive to our developing understanding of AI risks,** tracking both empirical evidence and technical trends. As models grow more powerful, and evaluations and incident reports demonstrate new risk areas, standards should recommend correspondingly greater caution. Standards-setting bodies should further develop the technical expertise to ”skate where the puck is going” and anticipate possible harms from growing AI capabilities and wider AI deployment.

3. **Governments should require audits of frontier AI systems during training and before deployment.**

    (a) **Governments should require AI developers to report efforts to create unusually or unpredictably capable AI systems.**

    (b) **Auditors and regulators should be granted the access necessary to evaluate models.** During training and before deployment of these frontier AI systems, governments should require that labs grant the access needed to evaluate these systems for dangerous capabilities. To protect labs' intellectual property, evaluation processes can use contractual and/or technical means, e.g. by using "structured access" methods rather than sharing model weights.

3. Governments should establish oversight of the AI industry and set consequences for AI harms.

To complement standards and auditing for reported AI systems, AI experts recommend that governments establish oversight and monitoring measures to track AI incidents and training runs. They further recommend establishing liability for AI harms, so as to incentivize responsible AI training and deployment.

1. **Governments should establish civil society oversight mechanisms for frontier AI development**. These include:

    (a) **Whistleblowing.** Governments should provide legal protections for whistleblowers at AI labs; employees at large tech companies, in particular, might otherwise fear retribution from their employers.

    (b) **Incident reports.** Governments should require labs to report incidents where AI systems displayed harmful behavior or dangerous capabilities.

2. **Governments should establish monitoring for large compute clusters.** They should monitor usage of government and industry supercomputers, including Know Your Customer (KYC) checks, and track the results in a central database.

3. **Governments should consolidate the above information in a registry of large AI systems** that are in training or deployment. This registry would track auditing results, incident reports, whistleblower disclosures, and compute usage, enabling regulators to identify potentially problematic systems.

4. **Governments should mandate that AI system developers and owners are legally liable** for harms from their AI systems that can be reasonably foreseen and prevented.

4. Governments should take further measures against emerging risks.

AI experts further call on governments to prepare standards and establish regulatory authorities which will be necessary for future AI systems with exceptionally dangerous capabilities.

1. **Governments should require information security measures for actors that will hold access to dangerous frontier AI systems, to prevent model proliferation.** Given the utility of advanced AI for economic gain and for malicious use, AI labs will need security measures of the highest standard, presenting a barrier even to Advanced Persistent Threats (APTs) and insider threats.

2. **Governments should prepare to establish a licensing system for training AI systems that may display highly dangerous capabilities**, such as resource-intensive frontier AI systems.

3. **Governments should empower regulators to pause the further development of an AI system that demonstrates dangerous capabilities during training.**

4. **Governments should mandate access controls for such frontier AI systems and their training code.** As suggested by Yoshua Bengio, one of the founders of deep learning, AI labs should limit external sharing of this information, and should keep employee access on a need-to-know basis.