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

In a recent short paper, world-leading AI scholars from the US, China, EU, UK, and other countries have highlighted that rapid AI progress will pose societal-scale risks. They urge their governments and leading AI labs to ensure responsible AI development. **This supplement highlights 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.**

Industry labs should invest in safe, ethical AI and develop responsible scaling plans.

Experts call on the leading AI labs to:

1. **Allocate at least one third of their AI R&D resources to ensure the safety and ethical use of AI systems.**[1] See below for suggested research topics.

2. **Promptly commit to detailed and independently scrutinized scaling policies**. These policies should describe specific safety measures that AI labs will take if specific dangerous capabilities are found in their AI systems. Scaling policies supplement regulation where it is too slow, yet cannot replace regulation.

Governments should invest in safe, ethical AI, establish oversight of the AI industry, and set consequences for AI harms.

Experts call on governments to:

1. **Allocate at least one third of their AI R&D resources**[1] **to ensure the safety and ethical use of AI systems**, comparable with their investment in AI capabilities. See below for suggested research topics.

2. **Establish oversight and monitoring of frontier AI**.

   (a) **Establish oversight of the AI industry by governments and civil society.**

      i. Provide legal protections for whistleblowers at major AI labs.
      ii. Create a registry of large AI systems that are in training or deployment.
      iii. Require labs to report incidents where AIs displayed harmful behavior or novel dangerous capabilities.

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

(b) **Monitor large compute clusters.**

i. Monitor usage of government and industry supercomputers and track results in a central database.

ii. Require compute providers to run Know Your Customer checks on clients performing AI training runs on large clusters.

3. **Require auditing of frontier AI systems during training and before deployment.**

   (a) Governments should mandate that AI labs report training runs above a particular computational budget — these "frontier AI systems" are likely to display unexpected properties.

   (b) During training and before deployment of these frontier AI systems, labs should give regulators and independent auditing bodies the access needed to evaluate these systems for dangerous capabilities. Evaluations can be performed without full model access, protecting AI labs' trade secrets.

4. **Create national and international safety standards that depend on model capabilities.** These standards should require more safety measures from AI labs and more careful auditing from third parties as AI systems grow more capable.

5. **Mandate that companies are legally liable for harms from their frontier AI systems that can be reasonably foreseen and prevented.**

6. **Establish or empower national institutions with the following desiderata:**

   (a) **Strong technical expertise.**

   (b) **Authority to facilitate international agreements and partnerships on AI governance.**

   (c) **Ability to set and enforce standards**, especially on high-risk frontier AI systems, while minimizing regulatory burden on AI systems that pose acceptable risk levels, allowing academic research to proceed.

   (d) **Authority to act swiftly** in the face of rapid AI progress.

Governments should take further measures against emerging risks.

The paper's authors further call on governments to establish standards and regulatory authorities for additional safety measures, which will be necessary for future AI systems with exceptionally dangerous capabilities.

1. **Establish a licensing system for training such AI systems** that are unusually resource-intensive and unusually risky.

2. **Empower regulators to pause the further development of an AI system**, if it demonstrates sufficiently dangerous capabilities during training.

3. **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.

4. **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, avoiding presenting a barrier even to Advanced Persistent Threats (APTs) and insider threats.

Details on R&D challenges

The authors highlight technical R&D areas to ensure the safety and ethical use of AI. These are not exhaustive. Quoting from the paper:

- 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.

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

- 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.

- Risk evaluations: Frontier AI systems develop unforeseen capabilities only discovered during training or even well after deployment. Better evaluation is needed to detect hazardous capabilities earlier.

- 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.

Additional relevant R&D areas are described in Hendrycks et al, cited by the authors. 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.