International Dialogue on AI Safety

Ditchley Park, UK
October 18-20, 2023

Statement

Coordinated global action on AI safety research and governance is critical to prevent uncontrolled frontier AI development from posing unacceptable risks to humanity.

Global action, cooperation, and capacity building are key to managing risk from AI and enabling humanity to share in its benefits. AI safety is a global public good that should be supported by public and private investment, with advances in safety shared widely. Governments around the world — especially of leading AI nations — have a responsibility to develop measures to prevent worst-case outcomes from malicious or careless actors and to rein in reckless competition. The international community should work to create an international coordination process for advanced AI in this vein.

We face near-term risks from malicious actors misusing frontier AI systems, with current safety filters integrated by developers easily bypassed. Frontier AI systems produce compelling misinformation and may soon be capable enough to help terrorists develop weapons of mass destruction. Moreover, there is a serious risk that future AI systems may escape human control altogether. Even aligned AI systems could destabilize or disempower existing institutions. Taken together, we believe AI may pose an existential risk to humanity in the coming decades.

In domestic regulation, we recommend mandatory registration for the creation, sale or use of models above a certain capability threshold, including open-source copies and derivatives, to enable governments to acquire critical and currently missing visibility into emerging risks. Governments should monitor large-scale data centers and track AI incidents, and should require that AI developers of frontier models be subject to independent third-party audits evaluating their information security and model safety. AI developers should also be required to share comprehensive risk assessments, policies around risk management, and predictions about their systems' behaviour in third party evaluations and post-deployment with relevant authorities.

We also recommend defining clear red lines that, if crossed, mandate immediate termination of an AI system — including all copies — through rapid and safe shut-down procedures. Governments should cooperate to instantiate and preserve this capacity. Moreover, prior to deployment as well as during training for the most advanced models, developers should demonstrate to regulators' satisfaction that their system(s) will not cross these red lines.

Reaching adequate safety levels for advanced AI will also require immense research progress. Advanced AI systems must be demonstrably aligned with their designer's intent, as well as appropriate norms and values. They must also be robust against both malicious actors and rare failure modes. Sufficient human control needs to be ensured for these systems. Concerted effort by the global research community in both AI and other disciplines is essential; we need a global network of dedicated AI safety research and governance institutions. We call on leading AI developers to make a minimum spending commitment of one third of their AI R&D on AI safety and for government agencies to fund academic and non-profit AI safety and governance research in at least the same proportion.

Signatories:
Andrew Yao, Dean, Institute for Interdisciplinary Information Sciences, Tsinghua University
Yoshua Bengio, Professor, Department of CS and Operations Research, Université de Montréal
Stuart Russell, Professor of Electrical Engineering and Computer Sciences, UC Berkeley
Ya-Qin Zhang, Dean, Institute for AI Industry Research, Tsinghua University
Ed Felten, Professor of Computer Science and Public Affairs, Princeton University
Roger Grosse, Associate Professor of Computer Science, University of Toronto
Gillian Hadfield, Professor of Law, University of Toronto
Sana Khareghani, Professor of Practice in AI, King's College London
Dylan Hadfield-Menell, Assistant Professor of Electrical Engineering and Computer Sciences, MIT
Karine Perset, Head, OECD.AI Policy Observatory
Dawn Song, Professor of Electrical Engineering and Computer Sciences, UC Berkeley
Xin Chen, PhD student, ETH Zurich
Max Tegmark, Professor of Physics, MIT
Elizabeth Seger, Research Scholar, Centre for the Governance of AI, Oxford
Yi Zeng, Professor, Institute of Automation, Chinese Academy of Sciences
HongJiang Zhang, Chairman, Beijing Academy of AI
Yang-Hui He, Fellow, London Institute
Adam Gleave, Founder and CEO, FAR Al
Fynn Heide, Research Scholar, Centre for the Governance of AI, Oxford