INSIGHTS

POLICY FORUM

ARTIFICIAL INTELLIGENCE

Managing extreme AI risks amid rapid progress

Preparation requires technical research and development, as well as adaptive, proactive governance

By **Yoshua Bengio[1], Geoffrey Hinton[2,3], Andrew Yao[4], Dawn Song[5], Pieter Abbeel[5], Trevor Darrell[5], Yuval Noah Harari[6], Ya-Qin Zhang[7], Lan Xue[8], Shai Shalev-Shwartz[9], Gillian Hadfield[3,10,11], Jeff Clune[3,12], Tegan Maharaj[3,11,13], Frank Hutter[14,15], Atılım Güneş Baydin[16], Sheila McIlraith[2,3,11], Qiqi Gao[17], Ashwin Acharya[18], David Krueger[19], Anca Dragan[5], Philip Torr[20], Stuart Russell[5], Daniel Kahneman[21], Jan Brauner[16,18], Sören Mindermann[1,16]**

Artificial intelligence (AI) is progressing rapidly, and companies are shifting their focus to developing generalist AI systems that can autonomously act and pursue goals. Increases in capabilities and autonomy may soon massively amplify AI's impact, with risks that include large-scale social harms, malicious uses, and an irreversible loss of human control over autonomous AI systems. Although researchers have warned of extreme risks from AI (*1*), there is a lack of consensus about how to manage them. Society's response, despite promising first steps, is incommensurate with the possibility of rapid, transformative progress that is expected by many experts. AI safety research is lagging. Present governance initiatives lack the mechanisms and institutions to prevent misuse and recklessness and barely address autonomous systems. Drawing on lessons learned from other safety-critical technologies, we outline a comprehensive plan that combines technical research and development (R&D) with proactive, adaptive governance mechanisms for a more commensurate preparation.

RAPID PROGRESS, HIGH STAKES

Present deep-learning systems still lack important capabilities, and we do not know how long it will take to develop them. However, companies are engaged in a race to create generalist AI systems that match or exceed human abilities in most cognitive work [see supplementary materials (SM)]. They are rapidly deploying resources and developing techniques to increase AI capabilities, with investment in training state-of-the-art models tripling annually (see SM).

There is much room for further advances because tech companies have the cash reserves needed to scale the latest training runs by multiples of 100 to 1000 (see SM). Hardware and algorithms will also improve: AI computing chips have been getting 1.4 times more cost-effective, and AI training algorithms 2.5 times more efficient, each year (see SM). Progress in AI also enables faster AI progress—AI assistants are increasingly used to automate programming, data collection, and chip design (see SM).

There is no fundamental reason for AI progress to slow or halt at human-level abilities. Indeed, AI has already surpassed human abilities in narrow domains such as playing strategy games and predicting how proteins fold (see SM). Compared with humans, AI systems can act faster, absorb more knowledge, and communicate at a higher bandwidth. Additionally, they can be scaled to use immense computational resources and can be replicated by the millions. We do not know for certain how the future of AI will unfold. However, we must take seriously the possibility that highly powerful generalist AI systems that outperform human abilities across many critical domains will be developed within this decade or the next. What happens then?

More capable AI systems have larger impacts. Especially as AI matches and surpasses human workers in capabilities and cost-effectiveness, we expect a massive increase in AI deployment, opportunities, and risks. If managed carefully and distributed fairly, AI could help humanity cure diseases, elevate living standards, and protect ecosystems. The opportunities are immense.

But alongside advanced AI capabilities come large-scale risks. AI systems threaten to amplify social injustice, erode social stability, enable large-scale criminal activity, and facilitate automated warfare, customized mass manipulation, and pervasive surveillance [(*2*); see SM].

Many risks could soon be amplified, and new risks created, as companies work to develop autonomous AI: systems that can use tools such as computers to act in the world and pursue goals (see SM). Malicious actors could deliberately embed undesirable goals. Without R&D breakthroughs (see next section), even well-meaning developers may inadvertently create AI systems that pursue unintended goals: The reward signal used to train AI systems usually fails to fully capture the intended objectives, leading to AI systems that pursue the literal specification rather than the intended outcome. Additionally, the training data never captures all relevant situations, leading to AI systems that pursue undesirable goals in new situations encountered after training.

Once autonomous AI systems pursue undesirable goals, we may be unable to keep them in check. Control of software is an old and unsolved problem: Computer worms have long been able to proliferate and avoid detection (see SM). However, AI is making progress in critical domains such as hacking, social manipulation, and strategic planning (see SM) and may soon pose unprecedented control challenges. To advance undesirable goals, AI systems could gain human trust, acquire resources, and influence key decision-makers. To avoid human intervention (*3*), they might copy their algorithms across global server networks (*4*). In open conflict, AI systems could autonomously deploy a variety of weapons, including biological ones. AI systems having access to such technology would merely continue existing trends to automate military activity. Finally, AI systems will not need to plot for influence if it is freely handed over. Companies, governments, and militaries may let autonomous AI systems assume critical societal roles in the name of efficiency.

Without sufficient caution, we may irreversibly lose control of autonomous AI systems, rendering human intervention ineffective. Large-scale cybercrime, social manipulation, and other harms could escalate rapidly. This unchecked AI advancement could culminate in a large-scale loss of life and the biosphere, and the marginalization or extinction of humanity.

We are not on track to handle these risks well. Humanity is pouring vast resources into making AI systems more powerful but far less into their safety and mitigating their harms. Only an estimated 1 to 3% of AI publications are on safety (see SM). For AI to be a boon, we must reorient; pushing AI capabilities alone is not enough.

We are already behind schedule for this reorientation. The scale of the risks means that we need to be proactive, because the

[1]Mila–Quebec AI Institute, Université de Montréal, Montreal, QC, Canada. [2]Department of Computer Science, University of Toronto, Toronto, ON, Canada. [3]Vector Institute, Toronto, ON, Canada. [4]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. [5]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA. [6]Department of History, The Hebrew University of Jerusalem, Jerusalem, Israel. [7]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China. [8]Institute for AI International Governance, Tsinghua University, Beijing, China. [9]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel. [10]Faculty of Law, University of Toronto, Toronto, ON, Canada. [11]Schwartz Reisman Institute for Technology and Society, University of Toronto, Toronto, ON, Canada. [12]Computer Science Department, University of British Columbia, Vancouver, BC, Canada. [13]Faculty of Information, University of Toronto, Toronto, ON, Canada. [14]ELLIS Institute Tübingen, Tübingen, Germany. [15]Department of Computer Science, University of Freiburg, Freiburg, Germany. [16]Department of Computer Science, University of Oxford, Oxford, UK. [17]Institute of Political Science, East China University of Political Science and Law, Shanghai, China. [18]RAND Corporation, Santa Monica, CA, USA. [19]Department of Engineering, University of Cambridge, Cambridge, UK. [20]Department of Engineering Science, University of Oxford, Oxford, UK. [21]School of Public and International Affairs, Princeton University, Princeton, NJ, USA. Email: jan.m.brauner@gmail.com

costs of being unprepared far outweigh those of premature preparation. We must anticipate the amplification of ongoing harms, as well as new risks, and prepare for the largest risks before they materialize.

REORIENT TECHNICAL R&D

There are many open technical challenges in ensuring the safety and ethical use of generalist, autonomous AI systems. Unlike advancing AI capabilities, these challenges cannot be addressed by simply using more computing power to train bigger models. They are unlikely to resolve automatically as AI systems get more capable [(*5*); see SM] and require dedicated research and engineering efforts. In some cases, leaps of progress may be needed; we thus do not know whether technical work can fundamentally solve these challenges in time. However, there has been comparatively little work on many of these challenges. More R&D may thus facilitate progress and reduce risks.

A first set of R&D areas needs breakthroughs to enable reliably safe AI. Without this progress, developers must either risk creating unsafe systems or falling behind competitors who are willing to take more risks. If ensuring safety remains too difficult, extreme governance measures would be needed to prevent corner-cutting driven by competition and overconfidence. These R&D challenges include the following:

Oversight and honesty More capable AI systems can better exploit weaknesses in technical oversight and testing, for example, by producing false but compelling output (see SM).

Robustness AI systems behave unpredictably in new situations. Whereas some aspects of robustness improve with model scale, other aspects do not or even get worse (see SM).

Interpretability and transparency AI decision-making is opaque, with larger, more capable models being more complex to interpret. So far, we can only test large models through trial and error. We need to learn to understand their inner workings (see SM).

Inclusive AI development AI advancement will need methods to mitigate biases and integrate the values of the many populations it will affect (see SM).

Addressing emerging challenges Future AI systems may exhibit failure modes that we have so far seen only in theory or lab experiments, such as AI systems taking control over the training reward-provision channels or exploiting weaknesses in our safety objectives and shutdown mechanisms to advance a particular goal (*3*, *6*–*8*). A second set of R&D challenges needs progress to enable effective, risk-adjusted governance or to reduce harms when safety and governance fail.

Evaluation for dangerous capabilities As AI developers scale their systems, unforeseen capabilities appear spontaneously, without explicit programming (see SM). They are often only discovered after deployment (see SM). We need rigorous methods to elicit and assess AI capabilities and to predict them before training. This includes both generic capabilities to achieve ambitious goals in the world (e.g., long-term planning and execution) as well as specific dangerous capabilities based on threat models (e.g., social manipulation or hacking). Present evaluations of frontier AI models for dangerous capabilities (*9*), which are key to various AI policy frameworks, are limited to spot-checks and attempted demonstrations in specific settings (see SM). These evaluations can sometimes demonstrate dangerous capabilities but cannot reliably rule them out: AI systems that lacked certain capabilities in the tests may well demonstrate them in slightly different settings or with posttraining enhancements. Decisions that depend on AI systems not crossing any red lines thus need large safety margins. Improved evaluation tools decrease the chance of missing dangerous capabilities, allowing for smaller margins.

Evaluating AI alignment If AI progress continues, AI systems will eventually possess highly dangerous capabilities. Before training and deploying such systems, we need methods to assess their propensity to use these capabilities. Purely behavioral evaluations may fail for advanced AI systems: Similar to humans, they might behave differently under evaluation, faking alignment (*6*–*8*).

Risk assessment We must learn to assess not just dangerous capabilities but also risk in a societal context, with complex interactions and vulnerabilities. Rigorous risk assessment for frontier AI systems remains an open challenge owing to their broad capabilities and pervasive deployment across diverse application areas (*10*).

Resilience Inevitably, some will misuse or act recklessly with AI. We need tools to detect and defend against AI-enabled threats such as large-scale influence operations, biological risks, and cyberattacks. However, as AI systems become more capable, they will eventually be able to circumvent human-made defenses. To enable more powerful AI-based defenses, we first need to learn how to make AI systems safe and aligned.

Given the stakes, we call on major tech companies and public funders to allocate at least one-third of their AI R&D budget, comparable to their funding for AI capabilities, toward addressing the above R&D challenges and ensuring AI safety and ethical use (*11*). Beyond traditional research grants, government support could include prizes, advance market commitments (see SM), and other incentives. Addressing these challenges, with an eye toward powerful future systems, must become central to our field.

GOVERNANCE MEASURES

We urgently need national institutions and international governance to enforce standards that prevent recklessness and misuse. Many areas of technology, from pharmaceuticals to financial systems and nuclear energy, show that society requires and effectively uses government oversight to reduce risks. However, governance frameworks for AI are far less developed and lag behind rapid technological progress. We can take inspiration from the governance of other safety-critical technologies while keeping the distinctiveness of advanced AI in mind—that it far outstrips other technologies in its potential to act and develop ideas autonomously, progress explosively, behave in an adversarial manner, and cause irreversible damage.

Governments worldwide have taken positive steps on frontier AI, with key players, including China, the United States, the European Union, and the United Kingdom, engaging in discussions and introducing initial guidelines or regulations (see SM). Despite their limitations—often voluntary adherence, limited geographic scope, and exclusion of high-risk areas like military and R&D-stage systems—these are important initial steps toward, among others, developer accountability, third-party audits, and industry standards.

Yet these governance plans fall critically short in view of the rapid progress in AI capabilities. We need governance measures that prepare us for sudden AI breakthroughs while being politically feasible despite disagreement and uncertainty about AI timelines. The key is policies that automatically trigger when AI hits certain capability milestones. If AI advances rapidly, strict requirements automatically take effect, but if progress slows, the requirements relax accordingly. Rapid, unpredictable progress also means that risk-reduction efforts must be proactive—identifying risks from next-generation systems and requiring developers to address them before taking high-risk actions. We

need fast-acting, tech-savvy institutions for AI oversight, mandatory and much-more rigorous risk assessments with enforceable consequences (including assessments that put the burden of proof on AI developers), and mitigation standards commensurate to powerful autonomous AI.

Without these, companies, militaries, and governments may seek a competitive edge by pushing AI capabilities to new heights while cutting corners on safety or by delegating key societal roles to autonomous AI systems with insufficient human oversight, reaping the rewards of AI development while leaving society to deal with the consequences.

Institutions to govern the rapidly moving frontier of AI To keep up with rapid progress and avoid quickly outdated, inflexible laws (see SM), national institutions need strong technical expertise and the authority to act swiftly. To facilitate technically demanding risk assessments and mitigations, they will require far greater funding and talent than they are due to receive under almost any present policy plan. To address international race dynamics, they need the affordance to facilitate international agreements and partnerships (see SM). Institutions should protect low-risk use and low-risk academic research by avoiding undue bureaucratic hurdles for small, predictable AI models. The most pressing scrutiny should be on AI systems at the frontier: the few most powerful systems, trained on billion-dollar supercomputers, that will have the most hazardous and unpredictable capabilities (see SM).

Government insight To identify risks, governments urgently need comprehensive insight into AI development. Regulators should mandate whistleblower protections, incident reporting, registration of key information on frontier AI systems and their datasets throughout their life cycle, and monitoring of model development and supercomputer usage (*12*). Recent policy developments should not stop at requiring that companies report the results of voluntary or underspecified model evaluations shortly before deployment (see SM). Regulators can and should require that frontier AI developers grant external auditors on-site, comprehensive ("white-box"), and fine-tuning access from the start of model development (see SM). This is needed to identify dangerous model capabilities such as autonomous self-replication, large-scale persuasion, breaking into computer systems, developing (autonomous) weapons, or making pandemic pathogens widely accessible [(*4*, *13*); see SM].

Safety cases Despite evaluations, we cannot consider coming powerful frontier AI systems "safe unless proven unsafe." With present testing methodologies, issues can easily be missed. Additionally, it is unclear whether governments can quickly build the immense expertise needed for reliable technical evaluations of AI capabilities and societal-scale risks. Given this, developers of frontier AI should carry the burden of proof to demonstrate that their plans keep risks within acceptable limits. By doing so, they would follow best practices for risk management from industries, such as aviation, medical devices, and defense software, in which companies make safety cases [(*14*, *15*); see SM]: structured arguments with falsifiable claims supported by evidence that identify potential hazards, describe mitigations, show that systems will not cross certain red lines, and model possible outcomes to assess risk. Safety cases could leverage developers' in-depth experience with their own systems. Safety cases are politically viable even when people disagree on how advanced AI will become because it is easier to demonstrate that a system is safe when its capabilities are limited. Governments are not passive recipients of safety cases: They set risk thresholds, codify best practices, employ experts and third-party auditors to assess safety cases and conduct independent model evaluations, and hold developers liable if their safety claims are later falsified.

Mitigation To keep AI risks within acceptable limits, we need governance mechanisms that are matched to the magnitude of the risks (see SM). Regulators should clarify legal responsibilities that arise from existing liability frameworks and hold frontier AI developers and owners legally accountable for harms from their models that can be reasonably foreseen and prevented, including harms that foreseeably arise from deploying powerful AI systems whose behavior they cannot predict. Liability, together with consequential evaluations and safety cases, can prevent harm and create much-needed incentives to invest in safety.

Commensurate mitigations are needed for exceptionally capable future AI systems, such as autonomous systems that could circumvent human control. Governments must be prepared to license their development, restrict their autonomy in key societal roles, halt their development and deployment in response to worrying capabilities, mandate access controls, and require information security measures robust to state-level hackers until adequate protections are ready. Governments should build these capacities now.

To bridge the time until regulations are complete, major AI companies should promptly lay out "if-then" commitments: specific safety measures they will take if specific red-line capabilities (*9*) are found in their AI systems. These commitments should be detailed and independently scrutinized. Regulators should encourage a race-to-the-top among companies by using the best-in-class commitments, together with other inputs, to inform standards that apply to all players.

To steer AI toward positive outcomes and away from catastrophe, we need to reorient. There is a responsible path—if we have the wisdom to take it. ■

REFERENCES AND NOTES

1. Center for AI Safety, Statement on AI risk (2023); https://www.safe.ai/work/statement-on-ai-risk.
2. L. Weidinger *et al*., in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2022), pp. 214–229.
3. D. Hadfield-Menell, A. Dragan, P. Abbeel, S. Russell, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, C. Sierra, Ed. (International Joint Conferences on Artificial Intelligence, 2017), pp. 220–227.
4. M. Kinniment *et al*., arXiv:2312.11671 (18 December 2023).
5. I. R. McKenzie *et al*., arXiv:2306.09479 (15 June 2023).
6. R. Ngo, L. Chan, S. Mindermann, arXiv:2209.00626 (20 August 2022).
7. E. Hubinger *et al*., arXiv:2401.05566 (10 January 2024).
8. M. K. Cohen, N. Kolt, Y. Bengio, G. K. Hadfield, S. Russell, *Science* **384**, 36 (2024).
9. T. Shevlane *et al*., arXiv:2305.15324 (24 May 2023).
10. L. Koessler, J. Schuett, arXiv:2307.08823 (17 July 2023).
11. D. Hendrycks, N. Carlini, J. Schulman, J. Steinhardt, arXiv:2109.13916 (28 September 2021).
12. N. Kolt *et al*., arXiv:2404.02675 (3 April 2024).
13. M. Phuong *et al*., arXiv:2403.13793 (20 March 2024).
14. J. Clymer, N. Gabrieli, D. Krueger, T. Larsen, arXiv:2403.10462 (15 March 2024).
15. T. A. Kelly, *SAE Trans. J. Mater. Manuf.* **113**, 257 (2004).

ACKNOWLEDGMENTS

J.B. and S.Mi. led this work and contributed equally to it. We dedicate this work with gratitude to the memory of Daniel Kahneman, our co-author, whose remarkable contributions to this paper and to humanity's cumulative knowledge and wisdom will never be forgotten. Y.B., J.C., G.Ha., and S.Mc. hold the position of Candian Institute for Advanced Research (CIFAR) AI Chair. J.C. is a senior research adviser to Google DeepMind. A.A. reports acting as an adviser to the Civic AI Security Program and was affiliated with the Institute for AI Policy and Strategy at the time of the first submission. A.D. now holds an appointment at Google DeepMind but joined the company after the manuscript was written. D.S. is the president of Oasis Labs. T.D. is a cofounder of Prompt AI. P.A. is a cofounder at covariant.ai and an investment partner at AIX Ventures. S.S.-S. is the chief technology officer at Mobileye. D.K. served as a research director for the UK Foundation Model Task Force in 2023 and joined the board of the nonprofit Center for AI Policy in 2024. G.Ha. reports the following activities: senior policy adviser at OpenAI from 2018 to 2023, member of the RAND Technology Advisory Group from 2023 to the present, and member of the Safety Critical AI Steering Committee of the Partnership on AI from 2022 to the present.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adn0117

Published online 20 May 2024
10.1126/science.adn0117