GlobalResearch
Center for Research on Globalization

“Large-scale Risks from Upcoming, Powerful AI Systems”: Managing AI Risks in an Era of Rapid Progress

By Yoshua Bengio, Geoffrey Hinton, Andrew Yao, and et al.
Global Research, October 25, 2023
Managing AI Risks 24 October 2023

***

Amid rapid AI progress, the authors of this paper express a consensus on the large-scale risks from upcoming, powerful AI systems. They call for urgent governance measures and a major shift in AI R&D towards safety and ethical practices before these systems are developed.

In 2019, GPT-2 could not reliably count to ten. Only four years later, deep learning systems can write software, generate photorealistic scenes on demand, advise on intellectual topics, and combine language and image processing to steer robots.

As AI developers scale these systems, unforeseen abilities and behaviors emerge spontaneously without explicit programming[1]. Progress in AI has been swift and, to many, surprising.

The pace of progress may surprise us again. Current 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[2, 3].

They are rapidly deploying more resources and developing new techniques to increase AI capabilities. Progress in AI also enables faster progress: AI assistants are increasingly used to automate programming [4] and data collection[5, 6] to further improve AI systems [7].

There is no fundamental reason why AI progress would slow or halt at the human level. Indeed, AI has already surpassed human abilities in narrow domains like protein folding or strategy games[8, 9, 10].

Compared to humans, AI systems can act faster, absorb more knowledge, and communicate at a far higher bandwidth. Additionally, they can be scaled to use immense computational resources and can be replicated by the millions.

The rate of improvement is already staggering, and tech companies have the cash reserves needed to scale the latest training runs by multiples of 100 to 1000 soon[11]. Combined with the ongoing growth and automation in AI R&D, we must take seriously the possibility that generalist AI systems will outperform human abilities across many critical domains within this decade or the next.

What happens then?

If managed carefully and distributed fairly, advanced AI systems could help humanity cure diseases, elevate living standards, and protect our ecosystems.

The opportunities AI offers are immense. But alongside advanced AI capabilities come large-scale risks that we are not on track to handle well. Humanity is pouring vast resources into making AI systems more powerful, but far less into safety and mitigating harms. For AI to be a boon, we must reorient; pushing AI capabilities alone is not enough.

We are already behind schedule for this reorientation. We must anticipate the amplification of ongoing harms, as well as novel risks, and prepare for the largest risks *well before they materialize*. Climate change has taken decades to be acknowledged and confronted; for AI, decades could be too long.

Societal-scale Risks

AI systems could rapidly come to outperform humans in an increasing number of tasks. If such systems are not carefully designed and deployed, they pose a range of societal-scale risks. They threaten to amplify social injustice, erode social stability, and weaken our shared understanding of reality that is foundational to society.

They could also enable large-scale criminal or terrorist activities.

Especially in the hands of a few powerful actors, AI could cement or exacerbate global inequities, or facilitate automated warfare, customized mass manipulation, and pervasive surveillance[12, 13].

Many of these risks could soon be amplified, and new risks created, as companies are developing *autonomous AI*: systems that can plan, act in the world, and pursue goals.

While current AI systems have limited autonomy, work is underway to change this[14]. For example, the non-autonomous GPT-4 model was quickly adapted to browse the web[15], design and execute chemistry experiments[16], and utilize software tools[17], including other AI models[18].

If we build highly advanced autonomous AI, we risk creating systems that pursue undesirable goals. Malicious actors could deliberately embed harmful objectives. Moreover, no one currently knows how to reliably align AI behavior with complex values. Even well-meaning developers may inadvertently build AI systems that pursue unintended goals—especially if, in a bid to win the AI race, they neglect expensive safety testing and human oversight.

Once autonomous AI systems pursue undesirable goals, embedded by malicious actors or by accident, 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[19]. However, AI is making progress in critical domains such as hacking, social manipulation, deception, and strategic planning[14, 20]. Advanced autonomous AI systems will pose unprecedented control challenges.

To advance undesirable goals, future autonomous AI systems could use undesirable strategies—learned from humans or developed independently—as a means to an end[21, 22, 23, 24]. AI systems could gain human trust, acquire financial resources, influence key decision-makers, and form coalitions with human actors and other AI systems.

To avoid human intervention[24], they could copy their algorithms across global server networks like computer worms. AI assistants are already co-writing a large share of computer code worldwide[25]; future AI systems could insert and then exploit security vulnerabilities to control the computer systems behind our communication, media, banking, supply-chains, militaries, and governments. In open conflict, AI systems could threaten with or use autonomous or biological weapons. AI having access to such technology would merely continue existing trends to automate military activity, biological research, and AI development itself. If AI systems pursued such strategies with sufficient skill, it would be difficult for humans to intervene.

Finally, AI systems may not need to plot for influence if it is freely handed over. As autonomous AI systems increasingly become faster and more cost-effective than human workers, a dilemma emerges.

Companies, governments, and militaries might be forced to deploy AI systems widely and cut back on expensive human verification of AI decisions, or risk being outcompeted[26, 27]. As a result, autonomous AI systems could increasingly assume critical societal roles.

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

Harms such as misinformation and discrimination from algorithms are already evident today[28]; other harms show signs of emerging[20]. It is vital to both address ongoing harms and anticipate emerging risks. This is *not* a question of either/or. Present and emerging risks often share similar mechanisms, patterns, and solutions[29]; investing in governance frameworks and AI safety will bear fruit on multiple fronts[30].

A Path Forward

If advanced autonomous AI systems were developed today, we would not know how to make them safe, nor how to properly test their safety. Even if we did, governments would lack the institutions to prevent misuse and uphold safe practices. That does not, however, mean there is no viable path forward. To ensure a positive outcome, we can and must pursue research breakthroughs in AI safety and ethics and promptly establish effective government oversight.

Reorienting Technical R&D

We need research breakthroughs to solve some of today's technical challenges in creating AI with safe and ethical objectives. Some of these challenges are unlikely to be solved by simply making AI systems more capable[22, 31, 32, 33, 34, 35]. These include:

- Oversight and honesty: More capable AI systems are better able to exploit weaknesses in oversight and testing[32, 36, 37]
  —for example, by producing false but compelling output[35, 38].
- Robustness: AI systems behave unpredictably in new situations (under distribution shift or adversarial inputs)[39, 40, 34].
- 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[41].
- Risk evaluations: Frontier AI systems develop unforeseen capabilities only discovered during training or even well after deployment[42]. Better evaluation is needed to detect hazardous capabilities earlier[43, 44].
- 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[24, 45].

Given the stakes, we call on major tech companies and public funders to allocate at least one-third of their AI R&D budget to ensuring safety and ethical use, comparable to their funding for AI capabilities. Addressing these problems[34], with an eye toward powerful future systems, must become central to our field.

Urgent Governance Measures

We urgently need national institutions and international governance to enforce standards in order to prevent recklessness and misuse. Many areas of technology, from pharmaceuticals to financial systems and nuclear energy, show that society both requires and effectively uses governance to reduce risks. However, no comparable governance frameworks are currently in place for AI.

Without them, companies and countries may seek a competitive edge by pushing AI capabilities to new heights while cutting corners on safety, or by delegating key societal roles to AI systems with little human oversight[26]. Like manufacturers releasing waste into rivers to cut costs, they may be tempted to reap the rewards of AI development while leaving society to deal with the consequences.

To keep up with rapid progress and avoid inflexible laws, national institutions need strong technical expertise and the authority to act swiftly. To address international race dynamics, they need the affordance to facilitate international agreements and partnerships[46, 47]. To protect low-risk use and academic research, they should avoid undue bureaucratic hurdles for small and predictable AI models. The most pressing scrutiny should be on AI systems at the frontier: a small number of most powerful AI systems – trained on billion-dollar supercomputers – which will have the most hazardous and unpredictable capabilities[48, 49].

To enable effective regulation, governments urgently need comprehensive insight into AI

development. Regulators should require model registration, whistleblower protections, incident reporting, and monitoring of model development and supercomputer usage[48, 50, 51, 52, 53, 54, 55]. Regulators also need access to advanced AI systems before deployment to evaluate them for dangerous capabilities such as autonomous self-replication, breaking into computer systems, or making pandemic pathogens widely accessible[43, 56, 57].

For AI systems with hazardous capabilities, we need a combination of governance mechanisms[48, 52, 58, 59] matched to the magnitude of their risks.

Regulators should create national and international safety standards that depend on model capabilities. They should also hold frontier AI developers and owners legally accountable for harms from their models that can be reasonably foreseen and prevented.

These measures can prevent harm and create much-needed incentives to invest in safety. Further measures are needed for exceptionally capable future AI systems, such as models that could circumvent human control.

Governments must be prepared to license their development, pause development in response to worrying capabilities, mandate access controls, and require information security measures robust to state-level hackers, until adequate protections are ready.To bridge the time until regulations are in place, major AI companies should promptly lay out if-then commitments: specific safety measures they will take if specific red-line capabilities are found in their AI systems. These commitments should be detailed and independently scrutinized.

AI may be the technology that shapes this century. While AI capabilities are advancing rapidly, progress in safety and governance is lagging behind. 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.

*

Note to readers: Please click the share button above. Follow us on Instagram and Twitter and subscribe to our Telegram Channel. Feel free to repost and share widely Global Research articles.

Authors

Yoshua Bengio, Mila – Quebec AI Institute, Université de Montréal, Canada CIFAR AI Chair

Geoffrey Hinton, University of Toronto, Vector Institute

Andrew Yao, Tsinghua University

Dawn Song, University of California, Berkeley

Pieter Abbeel, University of California, Berkeley

Yuval Noah Harari, The Hebrew University of Jerusalem, Department of History

Ya-Qin Zhang, Tsinghua University

Lan Xue, Tsinghua University, Institute for AI International Governance

Shai Shalev-Shwartz, The Hebrew University of Jerusalem

Gillian Hadfield, University of Toronto, SR Institute for Technology and Society, Vector Institute

Jeff Clune, University of British Columbia, Canada CIFAR AI Chair, Vector Institute

Tegan Maharaj, University of Toronto, Vector Institute

Frank Hutter, University of Freiburg

Atılım Güneş Baydin, University of Oxford

Sheila McIlrath, University of Toronto, Vector Institute

Qiqi Gao, East China University of Political Science and Law

Ashwin Acharya, Institute for AI Policy and Strategy

David Krueger, University of Cambridge

Anca Dragan, University of California, Berkeley

Philip Torr, University of Oxford

Stuart Russell, University of California, Berkeley

Daniel Kahneman, Princeton University, School of Public and International Affairs

Jan Brauner*, University of Oxford

Sören Mindermann*, Mila – Quebec AI Institute

Notes

1. Emergent Abilities of Large Language Models [link]
   Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S. and others,, 2022. Transactions on Machine Learning Research.
2. About [link]
   DeepMind,, 2023.
3. About [link]
   OpenAI,, 2023.
4. ML-Enhanced Code Completion Improves Developer Productivity [HTML]
   Tabachnyk, M., 2022. Google Research.
5. GPT-4 Technical Report [PDF]
   OpenAI,, 2023. arXiv [cs.CL].
6. Constitutional AI: Harmlessness from AI Feedback [PDF]
   Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A. and others,, 2022. arXiv [cs.CL].
7. Examples of AI Improving AI [link]
   Woodside, T. and Safety, C.f.A., 2023.

8. Highly Accurate Protein Structure Prediction with AlphaFold
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O. and others,, 2021. Nature, pp. 583–589.
9. Superhuman AI for Multiplayer Poker
Brown, N. and Sandholm, T., 2019. Science, pp. 885–890.
10. Deep Blue
Campbell, M., Hoane, A. and Hsu, F., 2002. Artificial Intelligence, pp. 57–83.
11. Alphabet Annual Report, page 33 [PDF]
Alphabet,, 2022.
12. An Overview of Catastrophic AI Risks [PDF]
Hendrycks, D., Mazeika, M. and Woodside, T., 2023. arXiv [cs.CY].
13. Taxonomy of Risks Posed by Language Models
Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P., Mellor, J. and others,, 2022. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 214–229.
14. A Survey on Large Language Model based Autonomous Agents [PDF]
Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J. and others,, 2023. arXiv [cs.AI].
15. ChatGPT plugins [link]
OpenAI,, 2023.
16. ChemCrow: Augmenting Large Language Models with Chemistry Tools [PDF]
Bran, A., Cox, S., White, A. and Schwaller, P., 2023. arXiv [physics.chem-ph].
17. Augmented Language Models: a Survey [PDF]
Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R. and others,, 2023. arXiv [cs.CL].
18. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face [PDF]
Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y. and others,, 2023. arXiv [cs.CL].
19. The Science of Computing: The Internet Worm
Denning, P., 1989. American Scientist, pp. 126–128.
20. AI Deception: A Survey of Examples, Risks, and Potential Solutions [PDF]
Park, P., Goldstein, S., O'Gara, A., Chen, M. and Hendrycks, D., 2023. arXiv [cs.CY].
21. Optimal Policies Tend to Seek Power [PDF]
Turner, A., Smith, L., Shah, R. and Critch, A., 2019. Thirty-Fifth Conference on Neural Information Processing Systems.
22. Discovering Language Model Behaviors with Model-Written Evaluations [PDF]
Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E. and Heiner, S., 2022. arXiv [cs.CL].
23. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark
Pan, A., Chan, J., Zou, A., Li, N., Basart, S. and Woodside, T., 2023. International Conference on Machine Learning.
24. The Off-Switch Game
Hadfield-Menell, D., Dragan, A., Abbeel, P. and Russell, S., 2017. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 220–227.
25. GitHub Copilot [link]
Dohmke, T., 2023.
26. Natural Selection Favors AIs over Humans [PDF]
Hendrycks, D., 2023. arXiv [cs.CY].
27. Harms from Increasingly Agentic Algorithmic Systems
Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N. and Krasheninnikov, D., 2023. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 651–666. Association for Computing Machinery.

28. On the Opportunities and Risks of Foundation Models [PDF]
Bommasani, R., Hudson, D., Adeli, E., Altman, R., Arora, S. and von Arx, S., 2021. arXiv [cs.LG].
29. AI Poses Doomsday Risks—But That Doesn't Mean We Shouldn't Talk About Present Harms Too [link]
Brauner, J. and Chan, A., 2023. Time.
30. Existing Policy Proposals Targeting Present and Future Harms [PDF]
Safety, C.f.A., 2023.
31. Inverse Scaling: When Bigger Isn't Better [PDF]
McKenzie, I., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A. and Prabhu, A., 2023. Transactions on Machine Learning Research.
32. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models [link]
Pan, A., Bhatia, K. and Steinhardt, J., 2022. International Conference on Learning Representations.
33. Simple Synthetic Data Reduces Sycophancy in Large Language Models [PDF]
Wei, J., Huang, D., Lu, Y., Zhou, D. and Le, Q., 2023. arXiv [cs.CL].
34. Unsolved Problems in ML Safety [PDF]
Hendrycks, D., Carlini, N., Schulman, J. and Steinhardt, J., 2021. arXiv [cs.LG].
35. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback [PDF]
Casper, S., Davies, X., Shi, C., Gilbert, T., Scheurer, J. and Rando, J., 2023. arXiv [cs.AI].
36. Consequences of Misaligned AI
Zhuang, S. and Hadfield-Menell, D., 2020. Advances in Neural Information Processing Systems, Vol 33, pp. 15763–15773.
37. Scaling Laws for Reward Model Overoptimization
Gao, L., Schulman, J. and Hilton, J., 2023. Proceedings of the 40th International Conference on Machine Learning, pp. 10835–10866. PMLR.
38. Learning from human preferences [link]
Amodei, D., Christiano, P. and Ray, A., 2017.
39. Goal Misgeneralization in Deep Reinforcement Learning [link]
Langosco di Langosco, A. and Chan, A., 2022. International Conference on Learning Representations.
40. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals [PDF]
Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J. and others,, 2022. arXiv [cs.LG].
41. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks
Räuker, T., Ho, A., Casper, S. and Hadfield-Menell, D., 2023. 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pp. 464–483.
42. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F. and others,, 2022. Advances in Neural Information Processing Systems, Vol 35, pp. 24824–24837.
43. Model evaluation for extreme risks [PDF]
Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J. and others,, 2023. arXiv [cs.AI].
44. Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries [PDF]
Koessler, L. and Schuett, J., 2023. arXiv [cs.CY].
45. The Alignment Problem from a Deep Learning Perspective [PDF]
Ngo, R., Chan, L. and Mindermann, S., 2022. arXiv [cs.AI].

46. International Institutions for Advanced AI
Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A. and others,, 2023. arXiv [cs.CY]. DOI: 10.48550/arXiv.2307.04699
47. International Governance of Civilian AI: A Jurisdictional Certification Approach [PDF]
Trager, R., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L. and others,, 2023.
48. Frontier AI Regulation: Managing Emerging Risks to Public Safety [PDF]
Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J. and others,, 2023. arXiv [cs.CY].
49. Predictability and Surprise in Large Generative Models
Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A. and others,, 2022. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1747–1764. Association for Computing Machinery.
50. It's Time to Create a National Registry for Large AI Models [link]
Hadfield, G., Cuéllar, M. and O'Reilly, T., 2023. Carnegie Endowment for International Piece.
51. Model Cards for Model Reporting
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B. and others,, 2019. FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 220–229.
52. General Purpose AI Poses Serious Risks, Should Not Be Excluded From the EU's AI Act | Policy Brief [link]
2023. AI Now Institute.
53. Artificial Intelligence Incident Database [link]
Database, A.I.I., 2023.
54. The Promise and Perils of Tech Whistleblowing [link]
Bloch-Wehba, H., 2023. Northwestern University Law Review, Forthcoming.
55. Proposing a Foundation Model Information-Sharing Regime for the UK [link]
Mulani, N. and Whittlestone, J., 2023. Centre for the Governance of AI.
56. Auditing Large Language Models: a Three-Layered Approach
Mökander, J., Schuett, J., Kirk, H. and Floridi, L., 2023. AI and Ethics. DOI: 10.1007/s43681-023-00289-2
57. Can Large Language Models Democratize Access to Dual-Use Biotechnology? [PDF]
Soice, E., Rocha, R., Cordova, K., Specter, M. and Esvelt, K., 2023. arXiv [cs.CY].
58. Towards Best Practices in AGI Safety and Governance: A survey of Expert Opinion [PDF]
Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E. and others,, 2023. arXiv [cs.CY].
59. Regulatory Markets: The Future of AI Governance [PDF]
Hadfield, G. and Clark, J., 2023. arXiv [cs.AI].

Featured image is from Pixabay

Articles by: Yoshua Bengio, Geoffrey Hinton, Andrew Yao, and et al.