Science
AAAS

Supplementary Materials for

Managing extreme AI risks amid rapid progress

Yoshua Bengio *et al.*

Corresponding author: Jan Brauner, jan.m.brauner@gmail.com

DOI: 10.1126/science.adn0117

The PDF file includes:

Supplementary References

Managing extreme AI risks amid rapid progress – Extended references

In this supplementary material, we provide a copy of the text with 73 additional citations, for readers who want to investigate the mentioned topics in more detail.

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]*

[1] Mila - Quebec AI Institute, Université de Montréal, Montreal, Canada
[2] Department of Computer Science, University of Toronto, Toronto, Canada
[3] Vector Institute, Toronto, Canada
[4] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China.
[5] Department of Electrical Engineering and Computer Science, University of California, Berkely, Berkeley, 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, Canada
[11] Schwartz Reisman Institute for Technology and Society, University of Toronto, Toronto, Canada
[12] Computer Science Department, University of British Columbia, Vancouver, Canada
[13] Faculty of Information, University of Toronto, Toronto, 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, 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, USA

Email: jan.m.brauner@gmail.com
*: equal contribution

[1–15]
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 [16,17]. They are rapidly deploying more resources and developing new techniques to increase AI capabilities, with investment in training state-of-the-art models tripling annually [18].

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 [19]. 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 [20,21]. Progress in AI also enables faster AI progress [22]—AI assistants are increasingly used to automate programming [23], data collection [24,25], and chip design [26].

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 [27–29]. 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 [2,30–34].

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 [35–39]. 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 [40]. However, AI is making progress in critical domains such as hacking, social manipulation, and strategic planning [35,41] 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 [42,43]. 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 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 well 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 [5,11,44–47] 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 [44,48,49], for example, by producing false but compelling output [45,50,51].

Robustness AI systems behave unpredictably in new situations. Whereas some aspects of robustness improve with model scale [52], other aspects do not or even get worse [11,53–55].

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 [56].

Inclusive AI development AI advancement will need methods to mitigate biases and integrate the values of the many populations it will affect [31,57].

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 [58]. They are often only discovered after deployment [59–61]. 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 [4,62,63]. 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 [64], 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 [65,66] and introducing initial guidelines or regulations [67–70]. 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 [71–73] 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 [74,75]. 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 [76,77].

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 [67,69]. 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 [78]. 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 [4,9,13,62,63,79].

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 [80], medical devices [81], and defense software [82], in which companies make safety cases [14,15,83–85]: 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 [76,86–88]. 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. Statement on AI Risk. 2023 [cited 1 May 2024]. Available: https://www.safe.ai/work/statement-on-ai-risk

2. Weidinger L, Uesato J, Rauh M, Griffin C, Huang P-S, Mellor J, et al. Taxonomy of Risks posed by Language Models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2022. pp. 214–229.

3. Hadfield-Menell D, Dragan A, Abbeel P, Russell S. The Off-Switch Game. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017; 220–227.

4. Kinniment M, Sato LJK, Du H, Goodrich B, Hasin M, Chan L, et al. Evaluating Language-Model Agents on Realistic Autonomous Tasks. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2312.11671

5. McKenzie IR, Lyzhov A, Pieler M, Parrish A, Mueller A, Prabhu A, et al. Inverse Scaling: When Bigger Isn't Better. Transactions on Machine Learning Research. 2023. Available: https://openreview.net/forum?id=DwgRm72GQF

6. Ngo R, Chan L, Mindermann S. The alignment problem from a deep learning perspective. International Conference on Learning Representations 2024. 2024. Available: https://openreview.net/forum?id=fh8EYKFKns

7. Hubinger E, Denison C, Mu J, Lambert M, Tong M, MacDiarmid M, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv [cs.CR]. 2024. Available: http://arxiv.org/abs/2401.05566

8. Cohen MK, Kolt N, Bengio Y, Hadfield GK, Russell S. Regulating advanced artificial agents. Science. 2024;384: 36–38.

9. Shevlane T, Farquhar S, Garfinkel B, Phuong M, Whittlestone J, Leung J, et al. Model evaluation for extreme risks. arXiv [cs.AI]. 2023. Available: http://arxiv.org/abs/2305.15324

10. Koessler L, Schuett J. Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries. arXiv [cs.CY]. 2023. Available: http://arxiv.org/abs/2307.08823

11. Hendrycks D, Carlini N, Schulman J, Steinhardt J. Unsolved Problems in ML Safety. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2109.13916

12. Kolt N, Anderljung M, Barnhart J, Brass A, Esvelt K, Hadfield GK, et al. Responsible Reporting for Frontier AI Development. arXiv [cs.CY]. 2024. Available: http://arxiv.org/abs/2404.02675

13. Phuong M, Aitchison M, Catt E, Cogan S, Kaskasoli A, Krakovna V, et al. Evaluating Frontier Models for Dangerous Capabilities. arXiv [cs.LG]. 2024. Available: http://arxiv.org/abs/2403.13793

14. Clymer J, Gabrieli N, Krueger D, Larsen T. Safety Cases: How to Justify the Safety of Advanced AI Systems. arXiv [cs.CY]. 2024. Available: http://arxiv.org/abs/2403.10462

15. Kelly T. A Systematic Approach to Safety Case Management. SAE Trans J Mater Manuf. 2004;113: 257–266.

16. DeepMind. About. [cited 15 Sep 2023]. Available: https://www.deepmind.com/about

17. OpenAI. About. [cited 15 Sep 2023]. Available: https://openai.com/about

18. Cottier B. Trends in the Dollar Training Cost of Machine Learning Systems. 2023. Available: https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems

19. Alphabet. Alphabet annual report, page 33 (page 71 in the pdf): “As of December 31, 2022, we had USD113.8 billion in cash, cash equivalents, and short-term marketable securities”. [For comparison, the cost of training GPT-4 has been estimated as USD50 million (https://epochai.org/trends), and Sam Altman, the CEO of OpenAI, has stated that the cost for the whole process was more than USD100 million (https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/).]. 2022. Available: https://abc.xyz/assets/d4/4f/a48b94d548d0b2fdc029a95e8c63/2022-alphabet-annual-report.pdf

20. Hobbhahn M, Heim L, Aydos G. Trends in Machine Learning Hardware. 2023. Available: https://epochai.org/blog/trends-in-machine-learning-hardware

21. Erdil E, Besiroglu T. Algorithmic progress in computer vision. arXiv [cs.CV]. 2022. Available: http://arxiv.org/abs/2212.05153

22. Examples of AI Improving AI. [cited 15 Sep 2023]. Available: https://ai-improving-ai.safe.ai/

23. Tabachnyk M. ML-Enhanced Code Completion Improves Developer Productivity. [cited 15 Sep 2023]. Available: https://blog.research.google/2022/07/ml-enhanced-code-completion-improves.html

24. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: Harmlessness from AI Feedback. arXiv [cs.CL]. 2022. Available: http://arxiv.org/abs/2212.08073

25. OpenAI. GPT-4 Technical Report. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.08774

26. Mirhoseini A, Goldie A, Yazgan M, Jiang JW, Songhori E, Wang S, et al. A graph placement methodology for fast chip design. Nature. 2021;594: 207–212.

27. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589.

28. Brown N, Sandholm T. Superhuman AI for multiplayer poker. Science. 2019;365: 885–890.

29. Campbell M, Hoane AJ, Hsu F-H. Deep Blue. Artificial Intelligence. 2002;134: 57–83.

30. Chan A, Salganik R, Markelius A, Pang C, Rajkumar N, Krasheninnikov D, et al. Harms from Increasingly Agentic Algorithmic Systems. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2023. pp. 651–666.

31. Eubanks V. Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor. St Martin’s Press; 2018.

32. Hendrycks D, Mazeika M, Woodside T. An Overview of Catastrophic AI Risks. arXiv [cs.CY]. 2023. Available: http://arxiv.org/abs/2306.12001

33. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the Opportunities and Risks of Foundation Models. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2108.07258

34. Solaiman I, Talat Z, Agnew W, Ahmad L, Baker D, Blodgett SL, et al. Evaluating the Social Impact of Generative AI Systems in Systems and Society. arXiv [cs.CY]. 2023. Available: http://arxiv.org/abs/2306.05949

35. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A Survey on Large Language Model based Autonomous Agents. arXiv [cs.AI]. 2023. Available: http://arxiv.org/abs/2308.11432

36. ChatGPT plugins. [cited 15 Sep 2023]. Available: https://openai.com/blog/chatgpt-plugins

37. Bran AM, Cox S, White AD, Schwaller P. ChemCrow: Augmenting large-language models with chemistry tools. arXiv [physics.chem-ph]. 2023. Available: http://arxiv.org/abs/2304.05376

38. Mialon G, Dessì R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, et al. Augmented Language Models: a Survey. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2302.07842

39. Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.17580

40. Denning PJ. The Science of Computing: The Internet Worm. American Scientist. 1989;77: 126–128.

41. Park PS, Goldstein S, O’Gara A, Chen M, Hendrycks D. AI Deception: A Survey of Examples, Risks, and Potential Solutions. arXiv [cs.CY]. 2023. Available: http://arxiv.org/abs/2308.14752

42. Toner H, Acharya A. Exploring Clusters of Research in Three Areas of AI Safety. 2022. Available: https://cset.georgetown.edu/publication/exploring-clusters-of-research-in-three-areas-of-ai-safety/

43. Emerging Technology Observatory. AI safety – ETO Research Almanac. [cited 12 Feb 2024]. Available: https://almanac.eto.tech/topics/ai-safety/

44. Pan A, Bhatia K, Steinhardt J. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. International Conference on Learning Representations. 2022 [cited 15 Sep 2023]. Available: https://openreview.net/forum?id=JYtwGwIL7ye

45. Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, Rando J, et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv [cs.AI]. 2023. Available: http://arxiv.org/abs/2307.15217

46. Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S, et al. Discovering Language Model Behaviors with Model-Written Evaluations. arXiv [cs.CL]. 2022. Available: http://arxiv.org/abs/2212.09251

47. Wei J, Huang D, Lu Y, Zhou D, Le QV. Simple synthetic data reduces sycophancy in large language models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2308.03958

48. Zhuang S, Hadfield-Menell D. Consequences of misaligned AI. Advances in Neural Information Processing Systems. 2020;33: 15763–15773.

49. Gao L, Schulman J, Hilton J. Scaling Laws for Reward Model Overoptimization. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th International Conference on Machine Learning. PMLR; 23--29 Jul 2023. pp. 10835–10866.

50. Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, et al. Towards Understanding Sycophancy in Language Models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2310.13548

51. Amodei D, Christiano P, Ray A. Learning from human preferences. [cited 15 Sep 2023]. Available: https://openai.com/research/learning-from-human-preferences

52. Hendrycks D, Basart S, Mu N, Kadavath S, Wang F, Dorundo E, et al. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. arXiv [cs.CV]. 2020. Available: http://arxiv.org/abs/2006.16241

53. Langosco LLD, Koch J, Sharkey LD, Pfau J, Krueger D. Goal Misgeneralization in Deep Reinforcement Learning. Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S, editors. 17--23 Jul 2022;162: 12004–12019.

54. Shah R, Varma V, Kumar R, Phuong M, Krakovna V, Uesato J, et al. Goal Misgeneralization: Why Correct Specifications Aren’t Enough For Correct Goals. arXiv [cs.LG]. 2022. Available: http://arxiv.org/abs/2210.01790

55. Wang TT, Gleave A, Tseng T, Pelrine K, Belrose N, Miller J, et al. Adversarial Policies Beat Superhuman Go AIs. arXiv [cs.LG]. 2022. Available: http://arxiv.org/abs/2211.00241

56. Räuker T, Ho A, Casper S, Hadfield-Menell D. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). 2023. pp. 464–483.

57. Sen A. Social Choice Theory. In: Arrow KJ, Intriligator M, editors. Handbook of Mathematical Economics, Vol III. Amsterdam: North Holland; 1986.

58. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models. Transactions on Machine Learning Research. 2022 [cited 15 Sep 2023]. Available: https://openreview.net/pdf?id=yzkSU5zdwD

59. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems. 2022;35: 24824–24837.

60. Zhou P, Pujara J, Ren X, Chen X, Cheng H-T, Le QV, et al. Self-Discover: Large Language Models Self-Compose Reasoning Structures. arXiv [cs.AI]. 2024. Available: http://arxiv.org/abs/2402.03620

61. Davidson T, Denain J-S, Villalobos P, Bas G. AI capabilities can be significantly improved without expensive retraining. arXiv [cs.AI]. 2023. Available: http://arxiv.org/abs/2312.07413

62. Mouton CA, Lucas C, Guest E. The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. Santa Monica, CA: RAND Corporation; 2024. doi:10.7249/RRA2977-2

63. Scheurer J, Balesni M, Hobbhahn M. Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2311.07590

64. Ho A, Taylor J. Using Advance Market Commitments for Public Purpose Technology Development. 2021. Available: https://www.belfercenter.org/publication/using-advance-market-commitments-public-purpose-technology-development

65. AI Safety Summit. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. 2023. Available: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

66. OECD. G7 Hiroshima Process on Generative Artificial Intelligence (AI). OECD Publishing; 2023. p. 37.

67. The White House (US). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 2023. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

68. Cyberspace Administration of China. Interim Measures for Generative Artificial Intelligence Service Management. 13 Jul 2023 [cited 12 Feb 2024]. Available: http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

69. European Union. EU AI Act. Jan 2024 [cited Feburary 12 2024]. Available: https://artificialintelligenceact.eu/the-act/

70. Department of State for Science, Innovation and Technology (UK). A pro-innovation approach to AI regulation. In: Gov.uk [Internet]. 29 Mar 2023 [cited 12 Feb 2024]. Available: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

71. Xue L, Jia K, Zhao J. Agile Governance Practices in Artificial Intelligence: Categorizing Regulatory Approaches and Constructing a Policy Toolbox. Chinese Public Administration. Forthcoming.

72. Maas MM. Aligning AI Regulation to Sociotechnical Change. In: Bullock JB, Chen Y-C, Himmelreich J, Hudson VM, Korinek A, Young MM, et al., editors. The Oxford Handbook of AI Governance. Oxford University Press;

73. Xue L, Zhao J. Toward Agile Governance: The Pattern of Emerging Industry Development and Regulation. Chinese Public Administration. 2019;410: 28–34.

74. Ho L, Barnhart J, Trager R, Bengio Y, Brundage M, Carnegie A, et al. International Institutions for Advanced AI. arXiv [cs.CY]. 2023. doi:10.48550/arXiv.2307.04699

75. Trager RF, Harack B, Reuel A, Carnegie A, Heim L, Ho L, et al. International Governance of Civilian AI: A Jurisdictional Certification Approach. Aug 2023. Available: https://cdn.governance.ai/International_Governance_of_Civilian_AI_OMS.pdf

76. Anderljung M, Barnhart J, Korinek A, Leung J, O'Keefe C, Whittlestone J, et al. Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv [cs.CY]. 2023. Available: http://arxiv.org/abs/2307.03718

77. Ganguli D, Hernandez D, Lovitt L, Askell A, Bai Y, Chen A, et al. Predictability and Surprise in Large Generative Models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2022. pp. 1747–1764.

78. Casper S, Ezell C, Siegmann C, Kolt N, Curtis TL, Bucknall B, et al. Black-Box Access is Insufficient for Rigorous AI Audits. arXiv [cs.CY]. 2024. Available: http://arxiv.org/abs/2401.14446

79. Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach. AI and Ethics. 2023. doi:10.1007/s43681-023-00289-2

80. European Organisation for the Safety of Air Navigation. EAD Safety Case Guidance. 2010. Available: https://www.eurocontrol.int/sites/default/files/2019-05/20101201-adq-ead-safety-case-guid-v1.0.pdf

81. Food and Drug Administration. Infusion Pumps Total Product Life Cycle - Guidance for Industry and FDA Staff. 2014. Available: https://www.fda.gov/media/78369/download

82. SMP12. Safety Case and Safety Case Report. Jun 2023 [cited 12 Feb 2024]. Available: https://www.asems.mod.uk/guidance/posms/smp12

83. Mcdermid J, Jia Y. Safety of artificial intelligence: A collaborative model. 2020. Available: https://ceur-ws.org/Vol-2640/paper_7.pdf

84. Iso/iec. ISO/IEC 23894:2023 Standard on Information technology — Artificial intelligence — Guidance on risk management. 2023.

85. Raz T, Hillson D. A Comparative Review of Risk Management Standards. Risk Manage: Int J. 2005;7: 53–66.

86. AI Now Institute. General Purpose AI Poses Serious Risks, Should Not Be Excluded From the EU's AI Act | Policy Brief. [cited 15 Sep 2023]. Available: https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act

87. Schuett J, Dreksler N, Anderljung M, McCaffary D, Heim L, Bluemke E, et al. Towards best practices in AGI safety and governance: A survey of expert opinion. arXiv [cs.CY]. 2023. Available: http://arxiv.org/abs/2305.07153

88. Hadfield GK, Clark J. Regulatory Markets: The Future of AI Governance. arXiv [cs.AI]. 2023. Available: http://arxiv.org/abs/2304.04914