Managing AI Risks in an Era of Rapid Progress

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	UC Berkeley
Pieter Abbeel	UC 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 McIlraith	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	UC Berkeley
Philip Torr	University of Oxford
Stuart Russell	UC Berkeley
Daniel Kahneman	Princeton University, School of Public and International Affairs
Jan Brauner*	University of Oxford
Sören Mindermann*	University of Oxford, Mila - Quebec AI Institute, Université de Montréal

Abstract

In this short consensus paper, we outline risks from upcoming, advanced AI systems. We examine large-scale social harms and malicious uses, as well as an irreversible loss of human control over autonomous AI systems. In light of rapid and continuing AI progress, we propose urgent priorities for AI R&D and governance.

Rapid AI progress

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].

*Equal contribution; correspondence to jan.m.brauner@gmail.com and soeren.mindermann@gmail.com.

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], data collection[5,6], and chip design[7] to improve AI systems further[8].

There is no fundamental reason why AI progress would slow or halt when it reaches human-level abilities. Indeed, AI has already surpassed human abilities in narrow domains like protein folding and strategy games[9–11]. 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[12]. 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 the current 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 various 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[13–16].

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[17]. For example, the non-autonomous GPT-4 model was quickly adapted to browse the web[18], design and execute chemistry experiments[19], and utilize software tools[20] including other AI models[21].

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; several research breakthroughs are needed (see below). 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[22]. However, AI is making progress in critical domains such as hacking, social manipulation, deception, and strategic planning[17,23]. 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[24–27]. 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[27], they might copy their algorithms across global server networks[28], as computer worms do. AI assistants are already co-writing a substantial share of computer code worldwide[29]; 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 systems 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 will 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[14,30]. 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[31]; other harms show signs of emerging[23]. 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[32]; investing in governance frameworks and AI safety will bear fruit on multiple fronts[33].

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[25,34–38]. These include:

- Oversight and honesty: More capable AI systems can better exploit weaknesses in oversight and testing[35,39,40]—for example, by producing false but compelling output[38,41,42].
- Robustness: AI systems behave unpredictably in new situations (under distribution shift or adversarial inputs)[37,43,44].
- Interpretability and transparency: 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[45].
- Inclusive AI development: AI advancement will need methods to mitigate biases and integrate the values of the many populations it will affect[15,46].
- Risk evaluations: Frontier AI systems develop unforeseen capabilities only discovered during training or even well after deployment[47]. Better evaluation is needed to detect hazardous capabilities earlier[48,49].
- 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[27,50].

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[37], 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 to prevent recklessness and misuse. Many areas of technology, from pharmaceuticals to financial systems and nuclear energy, show that society requires and effectively uses governance to reduce risks. However, no comparable governance frameworks are currently in place for AI. Without them, 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 AI systems with little human oversight. 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[51,52]. 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[53,54].

To enable effective regulation, governments urgently need comprehensive insight into AI development. Regulators should require registration of frontier systems in development, whistleblower protections, incident reporting, and monitoring of model development and supercomputer usage[53,55–60]. 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[48,61,62].

For AI systems with hazardous capabilities, we need a combination of governance mechanisms[53,57,63,64] matched to the magnitude of their risks. Regulators should create national and international safety standards that depend on model capabilities. These standards should follow best practices for risk management, including putting the onus on companies to show that their plans keep risks below an acceptable level[65]. 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 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[48] 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.

References

[1] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, Jun. 2022.

[2] DeepMind. "About." (n.d.), [Online]. Available: https://www.deepmind.com/about (visited on 09/15/2023).

[3] OpenAI. "About." (n.d.), [Online]. Available: https://openai.com/about (visited on 09/15/2023).

[4] M. Tabachnyk. "ML-Enhanced code completion improves developer productivity." (2022), [Online]. Available: https://blog.research.google/2022/07/ml-enhanced-code-completion-improves.html.

[5] OpenAI, "GPT-4 technical report," Mar. 2023. arXiv: 2303.08774 [cs.CL].

[6] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. El Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, "Constitutional AI: Harmlessness from AI feedback," Dec. 2022. arXiv: 2212.08073 [cs.CL].

[7] A. Mirhoseini, A. Goldie, M. Yazgan, J. W. Jiang, E. Songhori, S. Wang, Y.-J. Lee, E. Johnson, O. Pathak, A. Nazi, J. Pak, A. Tong, K. Srinivasa, W. Hang, E. Tuncer, Q. V. Le, J. Laudon, R. Ho, R. Carpenter, and J. Dean, "A graph placement methodology for fast chip design," *Nature*, vol. 594, no. 7862, pp. 207–212, Jun. 2021. DOI: 10.1038/s41586-021-03544-w.

[8] T. Woodside. "Examples of AI improving AI." (2023), [Online]. Available: https://ai-improving-ai.safe.ai/.

[9] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior,

K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021. DOI: 10.1038/s41586-021-03819-2.

[10] N. Brown and T. Sandholm, "Superhuman AI for multi-player poker," *Science*, vol. 365, no. 6456, pp. 885–890, Aug. 2019. DOI: 10.1126/science.aay2400.

[11] M. Campbell, A. J. Hoane, and F.-H. Hsu, "Deep blue," *Artificial Intelligence*, vol. 134, no. 1, pp. 57–83, Jan. 2002. DOI: 10.1016/S0004-3702(01)00129-1.

[12] 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.

[13] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel, "Taxonomy of risks posed by language models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, Seoul, Republic of Korea: Association for Computing Machinery, Jun. 2022, pp. 214–229. DOI: 10.1145/3531146.3533088.

[14] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, and T. Maharaj, "Harms from Increasingly Agentic Algorithmic Systems," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23, New York, NY, USA: Association for Computing Machinery, Jun. 12, 2023, pp. 651–666. DOI: 10.1145/3593013.3594033.

[15] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. St Martin's Press, 2018.

[16] D. Hendrycks, M. Mazeika, and T. Woodside, "An overview of catastrophic AI risks," Jun. 2023. arXiv: 2306.12001 [cs.CY].

[17] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen, "A survey on large language model based autonomous agents," Aug. 2023. arXiv: 2308.11432 [cs.AI].

[18] OpenAI. "ChatGPT plugins." (n.d.), [Online]. Available: https://openai.com/blog/chatgpt-plugins (visited on 10/16/2023).

[19] A. M. Bran, S. Cox, A. D. White, and P. Schwaller, "ChemCrow: Augmenting large-language models with chemistry tools," Apr. 2023. arXiv: 2304.05376 [physics.chem-ph].

[20] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, and T. Scialom, "Augmented language models: A survey," Feb. 2023. arXiv: 2302.07842 [cs.CL].

[21] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face," Mar. 2023. arXiv: 2303.17580 [cs.CL].

[22] P. J. Denning, "The science of computing: The internet worm," *American Scientist.*, vol. 77, no. 2, pp. 126–128, 1989.

[23] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks, "AI deception: A survey of examples, risks, and potential solutions," Aug. 2023. arXiv: 2308.14752 [cs.CY].

[24] A. M. Turner, L. R. Smith, R. Shah, A. Critch, and P. Tadepalli, "Optimal policies tend to seek power," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[25] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. El Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, and J. Kaplan, "Discovering language model behaviors with Model-Written evaluations," Dec. 2022. arXiv: 2212.09251 [cs.CL].

[26] A. Pan, J. S. Chan, A. Zou, N. Li, S. Basart, T. Woodside, J. Ng, H. Zhang, S. Emmons, and D. Hendrycks, "Do the rewards justify the means? measuring Trade-Offs between rewards and ethical behavior in the MACHIAVELLI benchmark," *International Conference on Machine Learning*, n.d.

[27] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "The Off-Switch game," *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 220–227, 2017.

[28] M. Kinniment, L. Jun, K. Sato, H. Du, B. Goodrich, M. Hasin, L. Chan, L. H. Miles, T. R. Lin, H. Wijk, J. Burget, A. Ho, E. Barnes, and P. Christiano, "Evaluating language-model agents on realistic autonomous tasks," 2023.

[29] T. Dohmke. "GitHub copilot." (n.d.), [Online]. Available: https://github.blog/2023-02-14-github-copilot-for-business-is-now-available/ (visited on 09/15/2023).

[30] A. Critch and S. Russell, "TASRA: A taxonomy and analysis of Societal-Scale risks from AI," Jun. 2023. arXiv: 2306.06924 [cs.AI].

[31] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," Center for Research on Foundation Models, Stanford University, 2021, https://crfm.stanford.edu/assets/report.pdf.

[32] J. Brauner and A. Chan, "AI poses doomsday Risks—But that doesn't mean we shouldn't talk about present harms too," *Time*, Aug. 2023.

[33] Center for AI Safety. "Existing policy proposals targeting present and future harms." (Jun. 2023), [Online]. Available: https://assets-global.website-files.com/63fe96aeda6bea77ac7d3000/647d5368c2368cc32b359f88%5C_Policy%5C%20Agreement%5C%20Statement.pdf (visited on 09/15/2023).

[34] I. R. McKenzie, A. Lyzhov, M. Pieler, A. Parrish, A. Mueller, A. Prabhu, E. McLean, A. Kirtland, A. Ross, A. Liu, A. Gritsevskiy, D. Wurgaft, D. Kauffman, G. Recchia, J. Liu, J. Cavanagh, M. Weiss, S. Huang, The Floating Droid, T. Tseng, T. Korbak, X. Shen, Y. Zhang, Z. Zhou, N. Kim, S. R. Bowman, and E. Perez, "Inverse scaling: When bigger isn't better," *Transactions on Machine Learning Research*, Oct. 2023, https://openreview.net/pdf?id=DwgRm72GQF.

[35] A. Pan, K. Bhatia, and J. Steinhardt, "The effects of reward misspecification: Mapping and mitigating misaligned models," in *International Conference on Learning Representations*, 2022.

[36] J. Wei, D. Huang, Y. Lu, D. Zhou, and Q. V. Le, "Simple synthetic data reduces sycophancy in large language models," Aug. 2023. arXiv: 2308.03958 [cs.CL].

[37] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, "Unsolved problems in ML safety," Sep. 2021. arXiv: 2109.13916 [cs.LG].

[38] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell, "Open problems and fundamental limitations of reinforcement learning from human feedback," Jul. 2023. arXiv: 2307.15217 [cs.AI].

[39] S. Zhuang and D. Hadfield-Menell, "Consequences of misaligned AI," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 763–15 773, 2020.

[40] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 10 835–10 866.

[41] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, *et al.*, "Towards understanding sycophancy in language models," *arXiv preprint arXiv:2310.13548*, 2023.

[42] D. Amodei, P. Christiano, and A. Ray. "Learning from human preferences," OpenAI. (Jun. 13, 2017), [Online]. Available: https://openai.com/research/learning-from-human-preferences (visited on 09/15/2023).

[43] L. L. D. Langosco, J. Koch, L. D. Sharkey, J. Pfau, and D. Krueger, "Goal misgeneralization in deep reinforcement learning," in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 2022, pp. 12 004–12 019.

[44] R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato, and Z. Kenton, "Goal misgeneralization: Why correct specifications aren't enough for correct goals," Oct. 2022. arXiv: 2210.01790 [cs.LG].

[45] T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell, "Toward transparent AI: A survey on interpreting the inner structures of deep neural networks," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, Feb. 2023, pp. 464–483. DOI: 10.1109/SaTML54575.2023.00039.

[46] A. Sen, "Social choice theory," in *Handbook of Mathematical Economics, Vol. III*, K. J. Arrow and M. Intriligator, Eds., Amsterdam: North Holland, 1986.

[47] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Jan. 2022. arXiv: 2201.11903 [cs.CL].

[48] T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano, and A. Dafoe, "Model evaluation for extreme risks," May 2023. arXiv: 2305.15324 [cs.AI].

[49] L. Koessler and J. Schuett, "Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries," Jul. 2023. arXiv: 2307.08823 [cs.CY].

[50] R. Ngo, L. Chan, and S. Mindermann, "The alignment problem from a deep learning perspective," Aug. 2022. arXiv: 2209.00626 [cs.AI].

[51] L. Ho, J. Barnhart, R. Trager, Y. Bengio, M. Brundage, A. Carnegie, R. Chowdhury, A. Dafoe, G. Hadfield, M. Levi, and D. Snidal, "International institutions for advanced AI," Jul. 2023. DOI: 10.48550/arXiv.2307.04699. arXiv: 2307.04699 [cs.CY].

[52] R. F. Trager, B. Harack, A. Reuel, A. Carnegie, L. Heim, L. Ho, S. Kreps, R. Lall, O. Larter, S. Ó hÉigeartaigh, S. Staffell, and J. J. Villalobos, "International goveranance of civilian AI: A jurisdictional certification approach," Oxford Martin AI Governance Initiative and Centre for the Governance of AI, Whitepaper, Aug. 2023.

[53] M. Anderljung, J. Barnhart, A. Korinek, J. Leung, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, B. Chang, T. Collins, T. Fist, G. Hadfield, A. Hayes, L. Ho, S. Hooker, E. Horvitz, N. Kolt, J. Schuett, Y. Shavit, D. Siddarth, R. Trager, and K. Wolf, "Frontier AI regulation: Managing emerging risks to public safety," Jul. 2023. arXiv: 2307.03718 [cs.CY].

[54] D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage, S. El Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, S. Johnston, A. Jones, N. Joseph, J. Kernian, S. Kravec, B. Mann, N. Nanda, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Kaplan, S. McCandlish, C. Olah, D. Amodei, and J. Clark, "Predictability and surprise in large generative models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, Seoul, Republic of Korea: Association for Computing Machinery, Jun. 2022, pp. 1747–1764. DOI: 10.1145/3531146.3533229.

[55] G. Hadfield, M. F. Cuéllar, and T. O'Reilly. "It's time to create a national registry for large AI models." (2023), [Online]. Available: https://carnegieendowment.org/2023/07/12/it-s-time-to-create-national-registry-for-large-ai-models-pub-90180.

[56] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, Atlanta, GA, USA: Association for Computing Machinery, Jan. 2019, pp. 220–229. DOI: 10.1145/3287560.3287596.

[57] AI Now Institute. "General purpose AI poses serious risks, should not be excluded from the EU's AI act — policy brief." (n.d.), [Online]. Available: https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act (visited on 09/15/2023).

[58] "Artificial intelligence incident database." (n.d.), [Online]. Available: https://incidentdatabase.ai/ (visited on 09/15/2023).

[59] H. Bloch-Wehba, "The Promise and Perils of Tech Whistleblowing," Texas A&M University School of Law, Research Paper 23-13, 2023.

[60] N. Mulani and J. Whittlestone. "Proposing a foundation model Information-Sharing regime for the UK." (n.d.), [Online]. Available: https://www.governance.ai/post/proposing-a-foundation-model-information-sharing-regime-for-the-uk (visited on 09/15/2023).

[61] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: A three-layered approach," *AI and Ethics*, May 2023. DOI: 10.1007/s43681-023-00289-2.

[62] E. H. Soice, R. Rocha, K. Cordova, M. Specter, and K. M. Esvelt, "Can large language models democratize access to dual-use biotechnology?," Jun. 2023. arXiv: 2306.03809 [cs.CY].

[63] J. Schuett, N. Dreksler, M. Anderljung, D. McCaffary, L. Heim, E. Bluemke, and B. Garfinkel, "Towards best practices in AGI safety and governance: A survey of expert opinion," May 2023. arXiv: 2305.07153 [cs.CY].

[64] G. K. Hadfield and J. Clark, "Regulatory Markets: The Future of AI Governance," Apr. 25, 2023. DOI: 10.48550/arXiv.2304.04914. arXiv: 2304.04914 [cs, econ, q-fin], preprint.

[65] Iso/iec, *ISO/IEC 23894:2023 standard on information technology — artificial intelligence — guidance on risk management*, Feb. 2023.