

5 Acknowledgements

Thanks to Katja Grace and Harlan Stewart in particular; to Michael Aird, Adam Bales, Rick Korzekwa, Fazl Barez, Sam Clark, Max Dalton, and many others for various levels of feedback and support; and to all the researchers we interviewed.

6 References

- AI Impacts. 2021. What do coherence arguments imply about the behavior of advanced AI? [https://wiki.aiimpacts.org/doku.php?id=agency:what_do_coherence_arguments_imply_about_the_behavior_of_advanced_ai&s\[\]=goal&s\[\]=directed](https://wiki.aiimpacts.org/doku.php?id=agency:what_do_coherence_arguments_imply_about_the_behavior_of_advanced_ai&s[]=goal&s[]=directed).
- AI Impacts. 2023a. Interview on the strength of the evidence for AI risk claims with an anonymous AI alignment researcher. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims/summary_of_an_interview_on_the_strength_of_the_evidence_for_ai_risk_claims_with_anonymous_ai_alignment_researcher.
- AI Impacts. 2023b. Interview with Jacob Hilton on the strength of the evidence for AI risk claims. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims/summary_of_an_interview_on_the_strength_of_the_evidence_for_ai_risk_claims_with_jacob_hilton. Online.
- AI Impacts. 2023c. Interview with Victoria Krakovna on the strength of the evidence for AI risk claims. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims/summary_of_an_interview_on_the_strength_of_the_evidence_for_ai_risk_claims_with_victoria_krakovna.
- AI Impacts. 2023d. Interviews on the strength of the evidence for AI risk claims. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims.
- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. <http://arxiv.org/abs/1606.06565>.
- Adam Bales. 2023. Will AI avoid exploitation? Artificial general intelligence and expected utility theory. *Philosophical Studies*.
- Nicholas Beale, Heather Battey, Anthony C. Davison, and Robert S. MacKay. 2020. An unethical optimization principle. *Royal Society Open Science*, 7(7).
- Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 2023. Taken out of context: On measuring situational awareness in LLMs. <http://arxiv.org/abs/2309.00667>.
- David C Berliner and Sharon L Nichols. 2005. The inevitable corruption of indicators and educators through high-stakes testing. Technical report, Arizona State University.
- Nick Bostrom. 2012. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2):71–85.
- Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, Oxford, United Kingdom.
- Oliver Braganza. 2022. Proxyeconomics, a theory and model of proxy-based competition and cultural evolution. *Royal Society Open Science*, 9(2).
- Donald T. Campbell. 1979. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90.

- Joseph Carlsmith. 2022. Is Power-Seeking AI an Existential Risk? <http://arxiv.org/abs/2206.13353>.
- Centre for AI Safety. 2023. Statement on AI Risk. <https://www.safe.ai/statement-on-ai-risk#open-letter>.
- Brian Christian. 2020. *The Alignment Problem – Machine Learning and Human Values*. W. W. Norton & Company, New York, NY.
- Alec Chrystal and Paul Mizen. 2003. Goodhart’s Law: its origins, meaning and implications for monetary policy. In *Central Banking, Monetary Theory and Practice: Essays in Honour of Charles Goodhart*. Edward Elgar Publishing.
- The Responsible AI Collective. 2023a. Incident 503: Bing AI Search Tool Reportedly Declared Threats against Users. <https://incidentdatabase.ai/cite/503/>.
- The Responsible AI Collective. 2023b. Incident 511: Microsoft’s Bing Failed to Fetch Movie Showtimes Results Due to Date Confusion. <https://incidentdatabase.ai/cite/511/>.
- Andrew Critch and David Krueger. 2020. *AI Research Considerations for Human Existential Safety (ARCHES)*. <http://arxiv.org/abs/2006.04948>.
- Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. 2021. *AI for radiographic COVID-19 detection selects shortcuts over signal*. *Nature Machine Intelligence*, 3(7):610–619.
- K Eric Drexler. 2019. Reframing Superintelligence. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf.
- EJT. 2023. There are no coherence theorems. <https://forum.effectivealtruism.org/posts/FoRyordtA7LDoEhd7/there-are-no-coherence-theorems>.
- Lewis Elton. 2004. *Goodhart’s Law and Performance Indicators in Higher Education*. *Evaluation & Research in Education*, 18(1-2):120–128.
- Michael Fire and Carlos Guestrin. 2019. *Over-optimization of academic publishing metrics: observing Goodhart’s Law in action*. *GigaScience*, 8(6).
- C. A. E. Goodhart. 1984. *Problems of Monetary Management: The UK Experience*. In *Monetary Theory and Practice: The UK Experience*. Macmillan Education UK.
- Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017. *The Off-Switch Game*. <http://arxiv.org/abs/1611.08219>.
- Dylan Hadfield-Menell and Gillian K. Hadfield. 2019. *Incomplete Contracting and AI Alignment*. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422. ACM.
- Rose Hadshar. 2023. Empirical evidence for existential AI risk factors. <https://airtable.com/embed/appWYvYLkBiDckhAo/shrkuKrEf4zhdVBrD/tbl3KurpJxkFVcNJJ?backgroundColor=red&viewControls=on>.
- Will Douglas Heaven. 2022. Why Meta’s latest large language model survived only three days online. <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/>.
- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021. *The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization*. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, Montreal, QC, Canada. IEEE.
- Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks. <http://arxiv.org/abs/2306.12001>.
- Christopher A. Hennessy and Charles A. E. Goodhart. 2023. *Goodhart’s Law and Machine Learning: A Structural Perspective*. *International Economic Review*, 64(3):1075–1086.

- Dario Amodei Jack Clark. 2016. Faulty reward functions in the wild. <https://openai.com/research/faulty-reward-functions>.
- Yohan J. John, Leigh Caldwell, Dakota E. McCoy, and Oliver Braganza. 2023. **Dead rats, dopamine, performance metrics, and peacock tails: proxy failure is an inherent risk in goal-oriented systems.** *Behavioral and Brain Sciences*, pages 1–68.
- Colm Kelly and Dennis J Snower. 2021. **Capitalism recoupled.** *Oxford Review of Economic Policy*, 37(4):851–863.
- Daniel Koretz. 2008. *Measuring Up: What Educational Testing Really Tells Us*. Harvard University Press.
- Victoria Krakovna and Janos Kramar. 2023. Power-seeking can be probable and predictive for trained agents. <http://arxiv.org/abs/2304.06528>.
- Viktoria Krakovna. 2020. Specification gaming examples in AI. <https://docs.google.com/spreadsheets/d/e/2PACX-1vRPipr0aC3HsCf5Tuum8bRfzYUiKLRqJmb0oC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bj0xCG84dAg/pubhtml>.
- Viktoria Krakovna, Jonathan Uesato, Vlad Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. 2020. Specification gaming: the flip side of AI ingenuity. <https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity>.
- Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. 2023. Goal Misgeneralization in Deep Reinforcement Learning. <http://arxiv.org/abs/2105.14111>.
- Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. 2017. AI Safety Gridworlds. <http://arxiv.org/abs/1711.09883>.
- Jiashuo Liu, Zheyang Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2023. Towards Out-Of-Distribution Generalization: A Survey. <http://arxiv.org/abs/2108.13624>.
- Robert E. Lucas. 1976. **Econometric policy evaluation: A critique.** *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46.
- David Manheim. 2019. **Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence.** *Big Data and Cognitive Computing*, 3(2).
- David Manheim and Scott Garrabrant. 2019. **Categorizing Variants of Goodhart’s Law.** <http://arxiv.org/abs/1803.04585>.
- Akhila Narla, Brett Kuprel, Kavita Sarin, Roberto Novoa, and Justin Ko. 2018. **Automated Classification of Skin Lesions: From Pixels to Practice.** *Journal of Investigative Dermatology*, 138(10):2108–2110.
- Richard Ngo, Lawrence Chan, and Sören Mindermann. 2023. The alignment problem from a deep learning perspective. <http://arxiv.org/abs/2209.00626>.
- Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. **Dissecting racial bias in an algorithm used to manage the health of populations.** *Science*, 366(6464):447–453.
- S O’Mahony. 2017. **Medicine and the Mcnamara Fallacy.** *Journal of the Royal College of Physicians of Edinburgh*, 47(3):281–287.
- Toby Ord. 2020. *The Precipice*. Bloomsbury Publishing.
- Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. **The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models.** <http://arxiv.org/abs/2201.03544>.

- Ethan Perez, Sam Ringer, Kamilė Lukošiuūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. *Discovering Language Model Behaviors with Model-Written Evaluations*. <http://arxiv.org/abs/2212.09251>.
- Michael Poku. 2016. *Campbell’s Law: implications for health care*. *Journal of Health Services Research & Policy*, 21(2):137–139.
- Salvador Pueyo. 2018. *Growth, degrowth, and the challenge of artificial superintelligence*. *Journal of Cleaner Production*, 197:1731–1736.
- Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2022. *Dataset Shift in Machine Learning*. MIT Press.
- Stuart Russell. 2019. *Human Compatible: AI and the Problem of Control*. Allen Lane, London.
- Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022a. Goal misgeneralization examples in AI. https://docs.google.com/spreadsheets/d/e/2PACX-1vTo3RkXUAigb25nP7gjpcHriR6XdzA_L5lo0cVFj_u7cRAZghWrYKH2L2nU4TA_Vr9KzBX5Bjz9G_1/pubhtml.
- Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022b. *Goal Misgeneralization: Why Correct Specifications Aren’t Enough For Correct Goals*. <http://arxiv.org/abs/2210.01790>.
- Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2023. Goal Misgeneralisation: Why Correct Specifications Aren’t Enough For Correct Goals. <https://deeplearningresearch.medium.com/goal-misgeneralisation-why-correct-specifications-arent-enough-for-correct-goals-cf96ebc60924>.
- Joar Skalse, Nikolaus Howe, Dmitrii Krashennnikov, and David Krueger. 2022. Defining and Characterizing Reward Gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.
- Marilyn Strathern. 1997. ‘Improving ratings’: audit in the British University system. *European Review*, 5(3):305–321.
- Wolfgang Stroebe. 2016. *Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations*. *Perspectives on Psychological Science*, 11(6):800–816.
- Rachel L. Thomas and David Uminsky. 2022. *Reliance on metrics is a fundamental challenge for AI*. *Patterns*, 3(5).
- Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2023. Optimal Policies Tend to Seek Power. <http://arxiv.org/abs/1912.01683>.
- Alexander Matt Turner and Prasad Tadepalli. 2022. *Parametrically Retargetable Decision-Makers Tend To Seek Power*. <http://arxiv.org/abs/2206.13477>.
- Innovation UK Parliament’s Science and Technology Committee. 2023. AI offers significant opportunities but twelve governance challenges must be addressed says science, innovation and technology committee. <https://committees.parliament.uk/committee/135/science-innovation-and-technology-committee/news/197236/ai-offers-significant-opportunities-but-twelve-governance-challenges-must-be-addressed-says-science-innovation-and-technology-committee/>.

Simon Zhuang and Dylan Hadfield-Menell. 2021. **Consequences of Misaligned AI**. <http://arxiv.org/abs/2102.03896>.

7 Appendix A: Carlsmith’s argument for existential risk via power-seeking AI

The following table maps between the premises of [Carlsmith \(2022\)](#)’s argument, and the claims used in this report (see Table 1). Claims within the scope of this report are bolded.

Note that the claims used in this report are not identical to Carlsmith’s premises, though they are closely related.

Carlsmith		Claims used in this report
By 2070:		<i>(Preconditions: Timelines)</i> The relevant AI systems will be developed in the not-too-distant future.
1. It will become possible and financially feasible to build AI systems with the following properties:	<i>Advanced capability:</i> they outperform the best humans on some set of tasks which when performed at advanced levels grant significant power in today’s world (tasks like scientific research, business/military/political strategy, engineering, and persuasion/manipulation).	<i>(Preconditions: Capabilities)</i> Some AI systems will be highly capable, in the sense that they are able to perform many important tasks at or above human level
	<i>Agentic planning:</i> they make and execute plans, in pursuit of objectives, on the basis of models of the world.	<i>(Preconditions: Goal-directedness)</i> Some AI systems will be goal-directed, in that they pursue goals consistently over long time periods.
	<i>Strategic awareness:</i> the models they use in making plans represent with reasonable accuracy the causal upshot of gaining and maintaining power over humans and the real-world environment.	<i>(Preconditions: Situational awareness)</i> Some AI systems will be aware that they are AI systems, and whether they are in training or deployment.
	(Call these “APS”—Advanced, Planning, Strategically aware—systems.)	
2. There will be strong incentives to build and deploy APS systems.		
3. It will be much harder to build APS systems that would not seek to gain and maintain power in unintended ways (because of problems with their objectives) on any of the inputs they’d encounter if deployed, than to build APS systems that would do this, but which are at least superficially attractive to deploy anyway.		<i>(Misalignment)</i> Some capable AI systems will develop goals which are misaligned with human goals. <i>(Misalignment: Specification gaming)</i> Some capable AI systems will learn designer-specified goals which diverge from intended goals in unforeseen ways. <i>(Misalignment: Goal misgeneralization)</i> Some capable AI systems will develop goals which are perfectly correlated with intended goals in training, but diverge once the systems are deployed.
4. Some deployed APS systems will be exposed to inputs where they seek power in unintended and high-impact ways (say, collectively causing >\$1 trillion dollars of damage), because of problems with their objectives.		<i>(Power-seeking)</i> Some capable, misaligned AI systems will seek power in order to achieve their goals.

5. Some of this power-seeking will scale (in aggregate) to the point of permanently disempowering all of humanity.	<i>(Existential consequences: Disempowerment)</i> This misaligned power-seeking will lead to permanent human disempowerment.
6. This disempowerment will constitute an existential catastrophe.	<i>(Existential consequences: Existential catastrophe)</i> Permanent human disempowerment will constitute an existential catastrophe.

8 Appendix B: Some evidence for other claims about existential risk from AI

We systematically reviewed the evidence for claims about misalignment and power-seeking. However, in the course of our research and interviews, we came across some evidence for other relevant claims.

This appendix contains some of the evidence for goal-directedness, situational awareness, and deceptive alignment. It should not be treated as a comprehensive review of the state of the evidence on these topics.

8.1 Some evidence for goal-directedness

Roughly, goal-directedness refers to a property of AI systems to persistently pursue a goal.³⁴ Goal-directedness has not been well-defined so far, and so reviewing the evidence for goal-directedness is hampered by unclarity about the concept.³⁵

That said, it seems plausible that goal-directedness is a direct precondition for goal misgeneralization and for power-seeking,³⁶ so it is an important claim to assess.

Coherence theorems offer one kind of conceptual evidence for goal-directedness, but the extent to which they apply to future AI systems is contested (Bales, 2023; EJT, 2023; AI Impacts, 2021).³⁷

There is limited empirical evidence of goal-directedness in systems so far.³⁸ One of the researchers we interviewed noted that language models may be particularly unsuited to goal-directedness.³⁹

However, individual researchers we interviewed believe that:

- To the extent that language models can simulate humans, they will have the ability to simulate goal-directedness.⁴⁰
- There is a clear trend towards systems acting more autonomously.⁴¹

³⁴In Carlsmith (2022), goal-directedness is referred to as “agentic planning”, where AI systems “make and execute plans, in pursuit of objectives, on the basis of models of the world.”

³⁵“Right now it’s really hard to distinguish between real goal-directedness and learned heuristics. . . I think part of the problem with goal-directedness is we don’t really understand the phenomenon that well.” [44:00] (AI Impacts, 2023c)

³⁶“Specifying something as goal misgeneralization also requires some assumption that the system is goal-directed to some degree and that can also be debatable.” [33:16] (AI Impacts, 2023c) “What I’m expecting is happening here is that current systems are not goal-directed enough to show real power-seeking. And so the power-seeking threat model becomes more reliant on these kind of extrapolations of when there are systems which are more capable, they’ll probably be at least somewhat more goal-directed and then once we have goal-directedness, we can more convincingly argue that power-seeking is going to be a thing because we have theory and so on, but there’s a lot of uncertainty about it because we don’t know how much systems will become more goal-directed.” [54:35] (AI Impacts, 2023c)

³⁷“Some of the theoretical arguments make the case that goal-directedness is an attractor. I think that’s something that’s more debatable, less clear to me. There have been various discussions on LessWrong and elsewhere about to what extent do coherence arguments imply goal-directedness. And I think the jury is still out on that one.” [42:36] (AI Impacts, 2023c)

³⁸“I think the evidence so far at least for language models, there isn’t really convincing evidence of goal-directedness.” [44:00] (AI Impacts, 2023c)

³⁹“It’s also possible goal-directedness is kind of hard. And especially, maybe language models are just a kind of system where goal-directedness comes less naturally than other systems like reinforcement learning systems or even with humans or whatever.” [40:26] (AI Impacts, 2023c)

⁴⁰“I think generally the kind of risk scenarios that we are most worried about would involve the system acting intentionally and deliberately towards some objectives but I would expect that intent and goal-directedness comes in degrees and if we see examples of increasing degrees of that then I think that does constitute evidence of that being possible. Although it’s not clear whether it will go all the way to really deliberate systems, but I think especially to the extent that these systems can simulate humans. . . they have the ability to simulate deliberate intentional action and planning because that’s something that humans can do.” [20:20] (AI Impacts, 2023c)

⁴¹“We are already capable of getting AI systems to do simple things relatively autonomously. I don’t think it’s a threshold where now it’s autonomous, now it’s not. . . I think it’s a spectrum and it’s just very clearly ramping up. We already have things that have a little autonomy but not very much. I think it’s just a pretty straightforward trend at this point.” [24:39] (AI Impacts, 2023b)

One researcher we interviewed highlighted goal-directedness as one of their key uncertainties about existential risk from AI.⁴²

8.2 Some evidence for situational awareness

“A model is situationally aware if it’s aware that it’s a model and can recognize whether it’s currently in testing or deployment.” (Berglund et al., 2023)

This is important to arguments about existential risk from AI as situational awareness is plausibly a precondition for successful misaligned power-seeking: a model may need to understand its own situation at a sophisticated level in order to make plans which successfully disempower humans. In particular, situational awareness seems like a precondition for deceptive alignment.

There is some empirical work demonstrating situational awareness in large language models, but the results are inconclusive (Berglund et al., 2023; Ngo et al., 2023; Perez et al., 2022). Berglund et al. (2023) find that language models can perform out-of-context reasoning tasks, but only with particular training set ups and data augmentation. Perez et al. (2022) run various experiments to test awareness, and find that “the models we evaluate are not aware of at least some basic details regarding themselves or their training procedures.” On the other hand, Langosco et al. (2023) use the same questions as Perez et al. (2022) but find that their model answers 85% accurately.

⁴²“I think we might see more goal-directed systems which produce clearer examples of internal goal misgeneralization, but also I wouldn’t be that surprised if we don’t see that. I think that’s one of the big uncertainties I have about level of risk. How much can we expect goal-directedness to emerge?” [40:26] (AI Impacts, 2023c)