INSTITUTE
FOR LAW & AI

Advanced AI governance

A literature review of problems, options, and proposals

AI Foundations Report 4 | Matthijs Maas | November 2023

law-ai.org

INSTITUTE
FOR LAW & AI

Advanced AI governance: A literature of problems, options, and proposals

Institute for Law & AI – AI Foundations Report 4

November 2023 | Matthijs Maas[1]

Abstract

As the capabilities of AI systems have continued to improve, the technology's global stakes have become increasingly clear. In response, an "advanced AI governance" community has come into its own, drawing on diverse bodies of research to analyze the potential problems this technology poses, map the options available for its governance, and articulate and advance concrete policy proposals. However, this field still faces a lack of internal and external clarity over its different research programmes. In response, this literature review provides an updated overview and taxonomy of research in advanced AI governance. After briefly setting out the aims, scope, and limits of this project, this review covers three major lines of work: (I) *problem-clarifying* research aimed at understanding the challenges advanced AI poses for governance, by mapping the strategic parameters (technical, deployment, governance) around its development and by deriving indirect guidance from history, models, or theory; (II) *option-identifying* work aimed at understanding affordances for governing these problems, by mapping potential key actors, their levers of governance over AI, and pathways to influence whether or how these are utilized; (III) *prescriptive* work aimed at identifying priorities and articulating concrete proposals for advanced AI policy, on the basis of certain views of the problem and governance options. The aim is that, by collecting and organizing the existing literature, this review will contribute to greater analytical and strategic clarity, enabling more focused and productive research, public debate, and policymaking on the critical challenges of advanced AI.

Cite as: Maas, Matthijs, 'Advanced AI governance: A literature review of problems, options, and proposals.' *Institute for Law & AI*, AI Foundations Report 4. (November 2023), https://law-ai.org/advanced-ai-gov-litrev/

[1] For comments and input on this document, I thank Di Cooke, Jeffrey Ding, Christoph Winter, Jonas Schuett, Oliver Guest, Gavin Leech, Peter Cihon, Jess Whittlestone, Eugenio Vargas Garcia, Jacob Arbeid, José Jaime Villalobos, Sam Clarke, Shin-Shin Hua, Richard Ngo, Liang Fang, Charlie Harrison, Leonie Koessler, Aishwarya Saxena, Moritz Kleinalterkamp, Carlos Ignacio Gutierrez, Charlotte Stix, Beth Barnes, Sascha Simon, Lewin Schmitt, Christian Ruhl, Zach Stein-Perlman, Lalitha Sundaram, Lara Thurnherr, and Anthony Barrett. I especially thank José Jaime Villalobos, Suzanne Van Arsdale, Alfredo Parra, Daisy Newbold-Harrop, Erin Cooper, and Wes Cowley for help in preparing this report for publication. All views expressed, and especially any remaining errors, are my own.

Executive Summary

This literature review provides an overview and taxonomy of past and recent research in the emerging field of advanced AI governance.

Aim: The aim of this review is to help disentangle and consolidate the field, improve its accessibility, enable clearer conversations and better evaluations, and contribute to overall strategic clarity or coherence in public and policy debates.

Summary: Accordingly, this review is organized as follows:

The introduction discusses the aims, scope, selection criteria, and limits of this review and provides a brief reading guide.

Part I reviews problem-clarifying work aimed at mapping the parameters of the AI governance challenge, including lines of research to map and understand:

1. Key technical parameters constituting the technical characteristics of advanced AI technology and its resulting (sociotechnical) impacts and risks. These include evaluations of the technical landscape of advanced AI (its forms, possible developmental pathways, timelines, trajectories), models for its general social impacts, threat models for potential extreme risks (based on general arguments and direct and indirect threat models), and the profile of the technical alignment problem and its dedicated research field.
2. Key deployment parameters constituting the conditions (present and future) of the AI development ecosystem and how these affect the distribution and disposition of the actors that will (first) deploy such systems. These include the size, productivity, and geographic distribution of the AI research field; key AI inputs; and the global AI supply chain.
3. Key governance parameters affecting the conditions (present and future) for governance interventions. These include stakeholder perceptions of AI and trust in its developers, the default regulatory landscape affecting AI, prevailing barriers to effective AI governance, and effects of AI systems on the tools of law and governance themselves.
4. Other lenses on characterizing the advanced AI governance problem. These include lessons derived from theory, from abstract models and wargames, from historical case studies (of technology development and proliferation, of its societal impacts and societal reactions, of successes and failures in historical attempts to initiate technology governance, and of successes and failures in the efficacy of different governance levers at regulating technology), and lessons derived from ethics and political theory.

Part II reviews option-identifying work aimed at mapping potential affordances and avenues for governance, including lines of research to map and understand:

1. Potential key actors shaping advanced AI, including actors such as or within AI labs and companies, the digital AI services and compute hardware supply chains, AI industry and academia, state and governmental actors (including the US, China, the EU, the UK, and other states), standard-setting organizations, international organizations, and public, civil society, and media actors.
2. Levers of governance available to each of these actors to shape AI directly or indirectly.
3. Pathways to influence on each of these key actors that may be available to (some) other actors in aiming to help inform or shape the key actors' decisions around whether or how to utilize key levers of governance to improve the governance of advanced AI.

Part III reviews prescriptive work aimed at putting this research into practice in order to improve the governance of advanced AI (for some view of the problem and of the options). This includes lines of research or advocacy to map, articulate, and advance:

1. Priorities for policy given theories of change based on some view of the problem and of the options.
2. Good heuristics for crafting AI policy. These include general heuristics for good regulation, for (international) institutional design, and for future-proofing governance.
3. Concrete policy proposals for the regulation of advanced AI, and the assets or products that can help these be realized and implemented. This includes proposals to regulate advanced AI using existing authorities, laws, or institutions; proposals to establish new policies, laws, or institutions (e.g., temporary or permanent pauses on AI development; the establishment of licensing regimes, lab-level safety practices, or governance regimes on AI inputs; new domestic governance institutions; new international AI research hubs; new bilateral agreements; new multilateral agreements; and new international governance institutions).

Table of contents

Introduction

This document aims to review, structure, and organize existing work in the field of advanced AI governance.

Background: Despite being a fairly young and interdisciplinary field, advanced AI governance offers a wealth of productive work to draw on and is increasingly structured through various research agendas[2] and syllabi.[3] However, while technical research on the possibility, impacts, and risks of advanced AI has been mapped in various literature reviews and distillations,[4] few attempts have been made to comprehensively map and integrate existing research on the governance of advanced AI.[5] This document aims to provide an overview and taxonomy of work in this field.

Aims: The aims of this review are several:

1. **Disentangle and consolidate the field** to promote greater clarity and legibility regarding the range of research, connections between different research streams and directions, and open gaps or underexplored questions. Literature reviews can contribute to such a consolidation of academic work;[6]

2. **Improve the field's accessibility and reduce some of its "research debt"**[7] to help those new to the field understand the existing literature, in order to facilitate a more cohesive and coordinated research field with lower barriers to entry, which reduces duplication of effort or work;

3. **Enable clearer conversations** between researchers exploring different questions or lines of research, discussing how and where their insights intersect or complement one another;

---

[2] For a number of influential research agendas, see: Dafoe, Allan. 'AI Governance: A Research Agenda'. Oxford: Center for the Governance of AI, Future of Humanity Institute, 2018. https://www.fhi.ox.ac.uk/govaiagenda/. Winter, Christoph, Jonas Schuett, Eric Martínez, Suzanne Van Arsdale, Renan Araújo, Nick Hollman, Jeff Sebo, Andrew Stawasz, Cullen O'Keefe, and Giuliana Rotola. 'Legal Priorities Research: A Research Agenda'. Legal Priorities Project, January 2021. https://www.legalpriorities.org/research_agenda.pdf. (Chapter 4); Clifton, Jesse. 'Cooperation, Conflict, and Transformative Artificial Intelligence- A Research Agenda'. Center on Long-Term Risk, March 2020. https://longtermrisk.org/files/Cooperation-Conflict-andTransformative-Artificial-Intelligence-A-Research-Agenda.pdf; Dafoe, Allan, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 'Open Problems in Cooperative AI'. ArXiv:2012.08630 [Cs], 15 December 2020. http://arxiv.org/abs/2012.08630; Gruetzemacher, Ross, Florian E. Dorner, Niko Bernaola-Alvarez, Charlie Giattino, and David Manheim. 'Forecasting AI Progress: A Research Agenda'. Technological Forecasting and Social Change 170 (1 September 2021): 120909. https://doi.org/10.1016/j.techfore.2021.120909.

[3] See also: Frazier, Kevin. 'Regulating AI: Legal and Policy Perspectives'. H2O, 2023. https://opencasebook.org/casebooks/9215/.; BlueDot Impact. 'AI Governance Curriculum'. AI Safety Fundamentals, 2022. https://aisafetyfundamentals.com/ai-governance-curriculum; and for older syllabi: Dafoe, Allan. 'Reading Guide for the Global Politics of Artificial Intelligence', 2017. https://www.allandafoe.com/aireadings.; Zwetsloot, Remco. 'Syllabus: Artificial Intelligence and International Security', 2018, 19. https://www.fhi.ox.ac.uk/wp-content/uploads/ArtificialIntelligence-and-International-Security-Syllabus.pdf

[4] See also the section below, on the profile of the technical alignment challenge. And also: Christiano, Paul. 'Current Work in AI Alignment'. Effective Altruism, 3 April 2020. https://www.effectivealtruism.org/articles/paul-christiano-current-work-in-ai-alignment.; for an older literature review of the technical AI risk field, see: Everitt, Tom, Gary Lea, and Marcus Hutter. 'AGI Safety Literature Review'. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018. http://arxiv.org/abs/1805.01109.

[5] For an excellent forthcoming introduction, see: Hendrycks, Dan, Thomas Woodside, Suryansh Mehta, Shankara Srikantan, Robert Trager, Jonas Schuett, Mauricio Baker, Lennart Heim, and Matthew Barnett. 'Governance'. In Introduction to AI Safety, Ethics, and Society, by Dan Hendrycks, 2023. https://www.aisafetybook.com/. And see previously: BlueDot Impact. 'AI Governance Curriculum'. AI Safety https://aisafetyfundamentals.com/ai-governance-curriculum.

[6] See broadly: Clancy, Matt. 'Literature Reviews and Innovation'. Substack newsletter. What's New the Sun (blog), 2 October 2023. https://mattsclancy.substack.com/p/literature-reviews-and-innovation?post_id=137592816&r=431 5a.

[7] Olah, Chris, and Shan Carter. 'Research Debt'. Distill 2, no. 3 (22 March 2017): e5. https://doi.org/10.23915/distill.00005.

4. **Enable better comparison** between different approaches and policy proposals; and

5. **Contribute to greater strategic clarity or coherence**,[8] improving the quality of interventions, and refining public and policy debates.

Scope: While there are many ways of framing the field, one approach is to define advanced AI governance as:

> **Advanced AI governance**: "the study and shaping of local and global governance systems—including norms, policies, laws, processes, and institutions—that affect the research, development, deployment, and use of existing and future AI systems, in ways that help the world choose the role of advanced AI systems in its future, and navigate the transition to that world."[9]

However, the aim of this document is not to engage in restrictive boundary policing of which research is part of this emerging field, let alone the "core" of it. The guiding heuristic here is not whether a given piece of research is directly, explicitly, and exclusively focused on certain "right" problems (e.g., extreme risks from advanced AI), nor whether it is motivated by certain political orientations or normative frameworks, nor even whether it explicitly uses certain terminology (e.g., "Transformative AI," "AGI," "General-Purpose AI System," or "Frontier AI").[10] Rather, the broad heuristic is simply whether the research helps answer a part of the advanced AI governance puzzle.

Accordingly, this review aims to cast a fairly broad net to cover work that meets any of the following criteria:

→ **Explicitly focuses on the governance of future advanced, potentially transformative AI systems**, in particular with regard to their potential significant impacts or extreme risks.

→ **Focuses on the governance of today's AI systems**, where (at least some of) the authors are interested in the implications of the analysis for the governance of future AI systems;

→ **Focuses on today's AI systems**, where the original work is (likely) not directly motivated by a concern over (risks from) advanced AI but nonetheless offers lessons that are or could be drawn upon by the advanced AI governance community to inform insights for the governance of advanced AI systems; and

→ **Focuses on (the impacts or governance of) non-AI technologies or issues** (such as historical case studies of technology governance), where the original work is not directly motivated by questions

[8] Notably, one might distinguish between (1) strategic clarity: achieving a sensible and grounded theory of change that provides a detailed model of the technical landscape and the policy world around advanced AI, with a resulting roadmap for how to select, evaluate, or prioritize present-day or near-term interventions; (2) strategic consensus: where (almost) everyone in a given epistemic community shares this same perspective or judgment; and (3) strategic coherence: when policy interventions or initiatives by different individuals or subcommunities in the field do not interfere with, counter, or erode one another (even if there remains underlying disagreement). Notably, while basic strategic clarity is invaluable for formulating robustly beneficial policies for advanced AI, it is unclear whether outright strategic consensus is always necessary or desirable, as a portfolio approach of many actors with different views (i.e., coherence, but lacking consensus) may be preferable. See: Maas, Matthijs. 'Components of Strategic Clarity'. EA Forum, 2 July 2022. https://forum.effectivealtruism.org/posts/ Bzezf2zmgBhtCD3Pb/components-of-strategic-clarity-strategic-perspectives-on.

[9] See also Maas, Matthijs, 'Concepts in Advanced AI Governance: A Literature Review of Key Terms and Definitions.' Institute for Law & AI. AI Foundations Report 3. (October 2023). https://www.legalpriorities.org/research/advanced-ai-gov-concepts Pg. 54. (discussing various technical, policy, and strategy-focused definitions of this field, and on that basis distilling this definition).

[10] Ibid

around AI but nonetheless offers lessons that are or could be drawn upon by the advanced AI governance community to inform insights for the governance of advanced AI systems.

Limitations: With this in mind, there are also a range of limitations or shortcomings for this review:

- → **Preliminary survey**: A literature review of this attempted breadth will inevitably fall short of covering all relevant work and sub-literatures in sufficient depth. In particular, given the speed of development in this field, a project like this will inevitably miss key work, so it should not be considered exhaustive. Indeed, because of the breadth of this report, I do not aim to go into the details of each topic, but rather to organize and list sources by topic. Likewise, there is some unbalance in that there has to date been more organized (technical) literature on (Part 1) characterizing the problem of advanced AI governance, than there has been on drafting concrete proposals (Part 3). As such, I invite others to produce "spin-offs" of this report which go into the detail of the content for each topic or sub-section in order to produce more in-depth literature reviews.[11]
- → **Broad scope**: In accordance with the above goal to cast a "broad net," this review covers both work that is core to and well established in the existing advanced AI governance field, and adjacent work that could be or has been considered by some as of significant value, even if it has not been as widely recognized yet. It also casts a broad net in terms of the type of sources surveyed, covering peer reviewed academic articles, reports, books, and more informal digital resources such as web fora.
- → **Incomplete in scope**: By and large, this review focuses on public and published analyses and mostly omits currently in-progress, unpublished, or draft work.[12] Given that a significant portion of relevant and key work in this field is unpublished, this means that this review likely will not capture all research directions in this field. Indeed, I estimate that this review captures at best ~70% of the work and research undertaken on many of these questions and subfields, and likely less. I therefore welcome further, focused literature reviews.
- → **A snapshot**: While this review covers a range of work, the field is highly dynamic and fast-moving, which means that this project will become outdated before long. Attempts will be made to update and reissue the report occasionally.

Finally, a few remaining disclaimers: (1) inclusion does not imply endorsement of a given article's conclusions; (2) this review aims to also highlight promising directions, such as issues or actors, that are not yet discussed in depth in the literature. As such, whenever I list certain issues (e.g., "actors" or "levers") without sources, this is because I have not yet found (or have missed out on) much work on that issue, suggesting there is a gap in the literature—and room for future work. Overall, this review should be seen as a living document that will be occasionally updated as the field develops. To that end, I welcome feedback, criticism, and suggestions for improvement.

Reading guide: In general, I recommend that rather than aiming to read this from the top, readers instead identify a theme or area of interest and jump to that section. In particular, this review may be most useful to readers (a) that already have a specific research question and want to see what work has been done and how a

[11] As one example of such a more targeted literature review, see also Maas, Matthijs M., and José Jaime Villalobos. 'International AI Institutions: A Literature Review of Models, Examples, and Proposals'. AI Foundations Report. Institute for Law & AI, September 2023. https://www.legalpriorities.org/research/international-ai-institutions.

[12] A small number of references to in-progress or forthcoming work are also included, with the authors' express consent.

particular line of work would fit into the larger landscape; (b) that aim to generate or distill syllabi for reading groups or courses; or (c) that aim to explore the broader landscape or build familiarity with fields or lines of research they have not previously explored. All the research presented here is collected from prior work, and I encourage readers to consult and directly cite those original sources named here.

I. Problem-clarifying work: Understanding the AI governance challenge

Most object-level work in the field of advanced AI governance has sought to disambiguate and reduce uncertainties around relevant strategic parameters of the AI governance challenge.[13]

> AI governance strategic parameters can be defined as "features of the world, such as the future AI development trajectory, the prevailing deployment landscape, and applicable policy conditions, which significantly determine the strategic nature of the advanced AI governance challenge."[14]

Strategic parameters serve as highly decision-relevant or even crucial considerations, determining which interventions or solutions are appropriate, necessary, viable, or beneficial for addressing the advanced AI governance challenge. Different views of these parameters constitute underlying cruxes for different theories of actions and approaches. This review discusses three types of strategic parameters:[15]

- → Technical parameters of the advanced AI challenge (i.e., what are the future technical developments in AI, on what timelines and on what trajectory will progress occur, why or how might such systems pose risks, and how difficult is the alignment challenge);
- → Deployment parameters of who is most likely to develop advanced AI systems and how they are likely to develop and use them (i.e., whose development decisions are to be governed); and
- → Governance parameters of how, when, and why governance interventions to shape advanced AI development and deployment are most likely to be viable, effective, or productive.

---

[13] Note, these are not exhaustive of all relevant key parameters for AI governance. For previous mappings of relevant (technical and governance) parameters for advanced AI governance, see also: Avin, Shahar. 'Exploring AGI Scenarios'. Presented at the FLI, 2019. https://futureoflife.org/wpcontent/uploads/2019/02/avin_friday_am.pdf?x76795.; Seth Baum on AI Governance, 2021. https://www.youtube.com/watch?v=G-8uEg7mCdA.; Hua, Shin-Shin, and Haydn Belfield. 'Effective Enforceability of EU Competition Law Under Different AI Development Scenarios: A Framework for Legal Analysis'. Verfassungsblog (blog), 18 August 2022. https://verfassungsblog.de/effective-enforceability-of-eu-competition-law-under-different-ai-develo pment-scenarios/. (sketching six "dimensions" for AI development scenarios—capability, speed of development, key inputs into AI development, model of AI system, number of actors, and nature of actor); Hobbhahn, Marius, Max Räuker, Yannick Mühlhäuser, Jasper Götting, and Simon Grimm. 'What Success Looks Like'. Effective Altruism Forum, 28 June 2022. https://forum.effectivealtruism.org/posts/AuRBKFnjABa6c6GzC/what-success-looks-like. (proposing a range of 'scenario variables'); Kilian, Kyle A., Christopher J. Ventura, and Mark M. Bailey. 'Examining the Differential Risk from High-Level Artificial Intelligence and the Question of Control'. Futures 151 (1 August 2023): 103182. https://doi.org/10.1016/j.futures.2023.103182. Pg. 7 (sketching 14 primary "dimensions" of AI technology transitions, which can see a total of 47 different "individual conditions" (future outcomes)).

[14] For discussion of these terms, see also Maas, Matthijs, "Disentangling Definitions in Advanced AI Governance'. Institute for Law & AI Foundations Report #2. Forthcoming 2023.

[15] See also Maas, Matthijs, 'Concepts in Advanced AI Governance: a Literature Review of Key Terms and Definitions.' Institute for Law & AI. AI Foundations Report #3. (October 2023). https://www.legalpriorities.org/research/advanced-ai-gov-concepts pg. 94. Others have referred to similar concepts by the term "(scenario) dimensions." See for instance: Hua, Shin-Shin, and Haydn Belfield. 'Effective Enforceability of EU Competition Law Under Different AI Development Scenarios: A Framework for Legal Analysis'. Verfassungsblog (blog), 18 August 2022. https://verfassungsblog.de/effective-enforceability-of-eu-competition-law-under-different-ai-develo pment-scenarios/.; Seth Baum on AI Governance 2021. https://www.youtube.com/watch?v=G-8uEg7mCdA. Kilian, Kyle A., Christopher J. Ventura, and Mark M. Bailey. 'Examining the Differential Risk from High-Level Artificial Intelligence and the Question Control'. Futures 151 https://doi.org/10.1016/j.futures.2023.103182. Pg. 7.

Accordingly, research in this subfield includes:

→ Empirical and theoretical work aiming to identify or get better estimates of each of these parameters as they apply to advanced AI (Sections 1, 2, 3).

→ Work applying other lenses to the advanced AI governance problem, drawing on other fields (existing theories, models, historical case studies, political and ethical theory) in order to derive crucial insights or actionable lessons (Section 4).

1. Technical parameters

An initial body of work focuses on mapping the relevant technical parameters of the challenge for advanced AI governance. This includes work on a range of topics relating to understanding the future technical landscape, understanding the likelihood of catastrophic risks given various specific threat models, and understanding the profile of the technical alignment problem and the prospects of it being solved by existing technical alignment research agendas.[16]

1.1. Advanced AI technical landscape

One subfield involves research to chart the future technical landscape of advanced AI systems.[17] Work to map this landscape includes research on the future form, pathways, timelines, and trajectories of advanced AI.

Forms of advanced AI

Work exploring distinct potential forms of advanced AI,[18] including:

→ strong AI,[19] autonomous machine intelligence,[20] general artificial intelligence,[21] human-level AI (HLAI),[22] general-purpose AI system (GPAIS),[23] comprehensive AI services (CAIS),[24] highly capable

[16] For an introduction to this field, see also: Hilton, Benjamin. 'Preventing an AI-Related Catastrophe Problem Profile'. 80,000 Hours, https://80000hours.org/problem-profiles/artificial-intelligence/.

[17] For a previous model and description of the "technical landscape," see Dafoe, Allan. 'AI Governance: A Research Agenda'. Oxford: Center for the Governance of AI, Future of Humanity Institute, 2018. https://www.fhi.ox.ac.uk/govaiagenda/. Pg. 15-33 (distinguishing the subfields "Mapping Technical Possibilities," "Assessing AI Progress," and "AI Safety").

[18] For a detailed survey of the range and varied definitions of each of these terms, see: Maas, Matthijs, 'Concepts in Advanced AI Governance: A Literature Review of Key Terms and Definitions.' Institute for Law & AI. AI Foundations Report 3. (October 2023). https://www.legalpriorities.org/research/advanced-ai-gov-concepts (Section II.1; and App. 1A).

[19] Searle, John R. 'Minds, Brains, and Programs'. Behavioral and Brain Sciences 3, no. 3 (September 1980): 417–24. https://doi.org/10.1017/S0140525X00005756. Pg. 417.; Russell, Stuart, and Peter Norvig. Artificial Intelligence: A Modern Approach. 3rd ed. Upper Saddle River: Pearson, 2016. Pg. 1020. Zeng, Yi, and Kang Sun. 'Whether We Can and Should Develop Strong AI: A Survey in China'. Center for Long-term Artificial Intelligence, 12 March 2023. https://long-term-ai.center/research/f/whether-we-can-and-should-develop-strong-artificial-intelligence.

[20] LeCun, Yann. 'A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27', 27 June 2022, 62. https://openreview.net/pdf?id=BZ5a1r-kVsf

[21] Hannas, William, Huey-Meei Chang, Daniel Chou, and Brian Fleeger. 'China's Advanced AI Research: Monitoring China's Paths to "General" Artificial Intelligence'. Center for Security and Emerging Technology, July 2022. https://cset.georgetown.edu/publication/chinas-advanced-ai-research/., pg. iii.

[22] McCarthy, John. 'From Here to Human-Level AI'. In Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning, 640–46. KR'96. Cambridge, Massachusetts, USA: Morgan Kaufmann Publishers Inc., 1996. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.384.8219&rep=rep1&type=pdf, pg 1175.; Nilsson, Nils J. 'Human-Level Artificial Intelligence? Be Serious!' AI Magazine, 2005. https://ai.stanford.edu/~nilsson/OnlinePubs-Nils/General%20Essays/AIMag26-04-HLAI.pdf; Muelhauser, Luke. 'What Is AGI?' Machine Intelligence Research Institute, 11 August 2013. https://intelligence.org/2013/08/11/what-is-agi/.; AI Impacts. 'Human-Level AI'. AI Impacts, 23 January 2014. https://aiimpacts.org/human-level-ai/. Shanahan, Murray. The Technological Singularity. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 229.

[23] Madiega, Tambiama. 'General-Purpose Artificial Intelligence'. EPRS (European Parliamentary Research Service), 2023. https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/745708/EPRS_ATA(2023)745708_EN.pdf. Pg. 1.; Gutierrez, Carlos I., Anthony Aguirre, Risto Uuk, Claire C. Boine, and Matija Franklin. 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems'. Digital Society 2, no. 3 (12 September 2023): 36. https://doi.org/10.1007/s44206-023-00068-w.

[24] Drexler, K Eric. 'Reframing Superintelligence: Comprehensive AI Services as General Intelligence'. Technical Report. Oxford: Future of Humanity Institute, University of Oxford, January 2019. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf. Pg. 1.

foundation models,[25] artificial general intelligence (AGI),[26] robust artificial intelligence,[27] AI+,[28] (machine/artificial) superintelligence,[29] and superhuman general purpose AI,[30] amongst others.

Developmental paths towards advanced AI

This includes research and debate on a range of domains. In particular, such work focuses on analyzing different hypothesized pathways towards achieving advanced AI based on different paradigms or theories.[31] Note that many of these are controversial and contested, and there is pervasive disagreement over the feasibility of many (or even all) of these approaches for producing advanced AI.

Nonetheless, some of these paradigms include programs to produce advanced AI based on:

[25] Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K Wei, Christoph Winter, et al. 'Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives'. Centre for the Governance of AI, 2023. https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models. Pg. 7.

[26] Adams, Sam, Itmar Arel, Joscha Bach, Robert Coop, Rod Furlan, Ben Goertzel, J. Storrs Hall, et al. 'Mapping the Landscape of Human-Level Artificial General Intelligence'. *AI Magazine* 33, no. 1 (15 March 2012): 25–42. https://doi.org/10.1609/aimag.v33i1.2322. Pg. 26.; Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 227.; Goertzel, Ben. 'Artificial General Intelligence: Concept, State of the Art, and Future Prospects'. *Journal of Artificial General Intelligence* 5, no. 1 (1 December 2014): 1–48. https://doi.org/10.2478/jagi-2014-0001. (pg 2); and see generally Goertzel, Ben, and Cassio Pennachin, eds. *Artificial General Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. https://doi.org/10.1007/978-3-540-68677-4_5.; Shevlin, Henry, Karina Vold, Matthew Crosby, and Marta Halina. 'The Limits of Machine Intelligence'. *EMBO Reports* 20, no. 10 (4 October 2019): e49177. https://doi.org/10.15252/embr.201949177.; Ngo, Richard. 'AGI Safety From First Principles', 2020. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ. Pg. 5. Muelhauser, Luke. 'What Is AGI?' Machine Intelligence Research Institute, 11 August 2013. https://intelligence.org/2013/08/11/what-is-agi/.; Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. Macmillan Publishers, 2019. https://us.macmillan.com/books/9780374715236/artificialintelligence.; ISO. 'ISO/IEC 22989:2022(En), Information Technology — Artificial Intelligence — Artificial Intelligence Concepts and Terminology'. Accessed 31 August 2023. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en.; Everitt, Tom, Gary Lea, and Marcus Hutter. 'AGI Safety Literature Review'. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 5441–49. IJCAI'18. Stockholm, Sweden: AAAI Press, 2018. https://dl.acm.org/doi/10.5555/3304652.3304782 pg. 5441.; Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. 'Sparks of Artificial General Intelligence: Early Experiments with GPT-4'. arXiv, 22 March 2023. https://doi.org/10.48550/arXiv.2303.12712. Pg. 4.

[27] Marcus, Gary. 'The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence'. arXiv, 19 February 2020. https://doi.org/10.48550/arXiv.2002.06177. Pg. 3.

[28] Chalmers, David J. 'The Singularity: A Philosophical Analysis'. *Journal of Consciousness Studies* 17 (2010): pg. 11.

[29] Bostrom, Nick. 'How Long Before Superintelligence?' *International Journal of Futures Studies* 2 (1998). https://nickbostrom.com/superintelligence. Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. Pg. 22. Barrett, Anthony M., and Seth D. Baum. 'A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis'. *Journal of Experimental & Theoretical Artificial Intelligence* 29, no. 2 (4 March 2017): 397–414. https://doi.org/10.1080/0952813X.2016.1186228. Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 231. Bostrom, Nick, Allan Dafoe, and Carrick Flynn. 'Public Policy and Superintelligent AI: A Vector Field Approach'. In *Ethics of Artificial Intelligence*, edited by S.M. Liao. Oxford University Press, 2019. http://www.nickbostrom.com/papers/aipolicy.pdf., pg 1–2.

[30] Aguirre, Anthony. 'Close the Gates to an Inhuman Future: How and Why We Should Choose to Not Develop Superhuman General-Purpose Artificial Intelligence'. SSRN Scholarly Paper. Rochester, NY, 20 October 2023. https://papers.ssrn.com/abstract=4608505. Pg. 1.

[31] For a detailed survey, see also: Maas, Matthijs, 'Concepts in Advanced AI Governance: A Literature Review of Key Terms and Definitions.' *Institute for Law & AI*. AI Foundations Report 3. (October 2023). https://www.legalpriorities.org/research/advanced-ai-gov-concepts (Section II.2; and Appendix 1B).

→ First principles: Approaches that aim to create advanced AI based on new fundamental insights in computer science, mathematics, algorithms, or software, producing AI systems that may, but need not, mimic human cognition.[32]

→ Direct/Scaling: Approaches that aim to "brute force" advanced AI[33] by running (one or more) existing AI approaches with increasingly greater computing power and/or training data to exploit observed "scaling laws" in system performance.[34]

→ Evolutionary: Approaches that aim to create advanced AI based on algorithms that compete to mimic the evolutionary brute search process that produced human intelligence.[35]

→ Reward-based: Approaches that aim to create advanced AI by running reinforcement learning systems with simple rewards in rich environments.[36]

→ Bootstrapping: Approaches that aim to create some minimally intelligent core system capable of subsequent recursive (self)-improvement as a "seed AI."[37]

→ Neuro-inspired: Various forms of biologically-inspired, brain-inspired, or brain-imitative approaches that aim to draw on neuroscience and/or "connectomics" to reproduce general intelligence.[38]

---

[32] Sotala, Kaj. 'Advantages of Artificial Intelligences, Uploads, and Digital Minds'. *International Journal of Machine Consciousness* 04, no. 01 (June 2012): 275–91. https://doi.org/10.1142/S1793843012400161. Pg. 1. ("AGI may be built on computer science principles and have little or no resemblance to the human psyche."); see also: Baum, Seth D., Ben Goertzel, and Ted G. Goertzel. 'How Long until Human-Level AI? Results from an Expert Assessment'. *Technological Forecasting and Social Change* 78, no. 1 (January 2011): 185–95. https://doi.org/10.1016/j.techfore.2010.09.006. pg. 19. ("many experts do not consider it likely that the first human-level AGI systems will closely mimic human intelligence").

[33] Hammond, Samuel. 'Why AGI Is Closer than You Think'. Second Best, 22 September 2023. https://www.secondbest.ca/p/why-agi-is-closer-than-you-think.

[34] See generally Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 'Scaling Laws for Neural Language Models'. *ArXiv:2001.08361 [Cs, Stat]*, 22 January 2020. http://arxiv.org/abs/2001.08361. See also Villalobos, Pablo. 'Scaling Laws Literature Review'. Epoch, 26 January 2023. https://epochai.org/blog/scaling-laws-literature-review.

[35] Carl Shulman and Nick Bostrom, "How Hard Is Artificial Intelligence? Evolutionary Arguments and Selection Effects," Journal of Consciousness Studies 19.7-8, 2012. https://nickbostrom.com/aievolution.pdf Note, this is distinct from the argument that evolutionary competitive pressures among human organizations (developing AI) may shape the development landscape for successful AI systems, especially in ways that promote the development of advanced AI agents with undesirable traits. See: Hendrycks, Dan. 'Natural Selection Favors AIs over Humans'. arXiv, 28 March 2023. https://doi.org/10.48550/arXiv.2303.16200.

[36] Silver, David, Satinder Singh, Doina Precup, and Richard S. Sutton. 'Reward Is Enough'. *Artificial Intelligence* 299 (1 October 2021): 103535. https://doi.org/10.1016/j.artint.2021.103535.

[37] Hall, John Storrs. 'Self-Improving AI: An Analysis'. *Minds and Machines* 17, no. 3 (1 October 2007): 249–59. https://doi.org/10.1007/s11023-007-9065-3.; Yudkowsky, Eliezer. 'Levels of Organization in General Intelligence'. In *Artificial General Intelligence*, edited by Ben Goertzel and Cassio Pennachin, 389–501. Cognitive Technologies. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. https://doi.org/10.1007/978-3-540-68677-4_12 Pg. 96. See also Yudkowsky, Eliezer. 'General Intelligence and Seed AI'. Singularity Institute, 2001. https://web.archive.org/web/20120805130100/singularity.org/files/GISAI.html. Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 230.

[38] See for instance: Zador, Anthony, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, et al. 'Catalyzing Next-Generation Artificial Intelligence through NeuroAI'. *Nature Communications* 14, no. 1 (22 March 2023): 1597. Pg. 2. https://doi.org/10.1038/s41467-023-37180-x; Eth, Daniel. 'The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes'. *Informatica* 41, no. 4 (27 December 2017). http://www.informatica.si/index.php/informatica/article/view/1874. See also: Farisco, Michele, Gianluca Baldassarre, Emilio Cartoni, Antonia Leach, Mihai A. Petrovici, Achim Rosemann, Arleen Salles, Bernd Stahl, and Sacha J. van Albada. 'A Method for the Ethical Analysis of Brain-Inspired AI'. arXiv, 18 May 2023. https://doi.org/10.48550/arXiv.2305.10938, pg. 4. I thank Carla Zoe Cremer for this suggestion.

→ Neuro-emulated: Approaches that aim to digitally simulate or recreate the states of human brains at a fine-grained level, possibly producing whole-brain-emulation.[39]

→ Neuro-integrationist: Approaches that aim to create advanced AI based on merging components of human and digital cognition.

→ Embodiment: Approaches that aim to create advanced AI by providing the AI system with a robotic physical "body"' to ground cognition and enable it to learn from direct experience of the world.[40]

→ Hybrid: Approaches that rely on combining deep neural network-based approaches to AI with other paradigms (such as symbolic AI).[41]

Notably, of these approaches, recent years have seen most sustained attention focused on the direct (scaling) approach and whether current approaches to advanced AI, if scaled up with enough computing power or training data, will suffice to produce advanced or transformative AI capabilities. There have been various arguments both in favor of and against this direct path.

→ Arguments in favor of a direct path: "scaling hypothesis,"[42] "prosaic AGI,"[43] and "Human feedback on diverse tasks (HFDT)";[44]

[39] Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 232.; Bostrom, Nick, and Anders Sandberg. 'Whole Brain Emulation: A Roadmap'. Technical Report. Future of Humanity Institute, 2008. http://www.fhi.ox.ac.uk/reports/2008-3.pdf. Pg 7.

[40] Gopalakrishnan, Keerthana. 'Embodiment Is Indispensable for AGI', 7 June 2022. https://keerthanapg.com/tech/embodiment-agi/ or https://www.lesswrong.com/posts/vBBxKBWn4zRXwivxC/embodiment-is-indispensable-for-agi ; Kremelberg, David. 'Embodiment as a Necessary a Priori of General Intelligence'. In *Artificial General Intelligence*, edited by Patrick Hammer, Pulin Agrawal, Ben Goertzel, and Matthew Iklé, 11654:132–36. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-27005-6_13.

[41] See for instance "hybrid AI": Marcus, Gary. 'The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence'. arXiv, 19 February 2020. https://doi.org/10.48550/arXiv.2002.06177.

[42] Among others, see famously: Sutton, Rich. 'The Bitter Lesson'. *Incomplete Ideas* (blog), 2019. http://www.incompleteideas.net/IncIdeas/BitterLesson.html. For an articulation of this 'scaling hypothesis', see also: Branwen, Gwern. 'The Scaling Hypothesis', 28 May 2020. https://www.gwern.net/Scaling-hypothesis.

[43] Christiano, Paul. 'Prosaic AI Alignment'. Medium, 28 March 2017. https://ai-alignment.com/prosaic-ai-control-b959644d79c2.

[44] Cotra, Ajeya. 'Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover'. AI Alignment Forum, 18 July 2022. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.

→ Arguments against a direct path, highlighting various limits and barriers: "deep limitations,"[45] "the limits of machine intelligence,"[46] "why AI is harder than we think,"[47] and other skeptical arguments;[48]

→ Discussion of the possible features of "engineering roadmaps" for AGI-like systems.[49]

Advanced AI timelines: Approaches and lines of evidence

A core aim of the field is to chart the timelines for advanced AI development across the future technical development landscape.[50] This research focuses on various lines of evidence,[51] which are here listed in order from more abstract to more concrete and empirical, and from relying more on outside-view arguments to

[45] Cremer, Carla Zoe. 'Deep Limitations? Examining Expert Disagreement over Deep Learning'. *Progress in Artificial Intelligence*, 26 June 2021. https://doi.org/10.1007/s13748-021-00239-1.

[46] Shevlin, Henry, Karina Vold, Matthew Crosby, and Marta Halina. 'The Limits of Machine Intelligence'. *EMBO Reports* 20, no. 10 (4 October 2019): e49177. https://doi.org/10.15252/embr.201949177.

[47] Mitchell, Melanie. 'Why AI Is Harder Than We Think'. *ArXiv:2104.12871 [Cs]*, 26 April 2021. http://arxiv.org/abs/2104.12871.

[48] Long, Robert, and Asya Bergal. 'Evidence against Current Methods Leading to Human Level Artificial Intelligence'. *AI Impacts* (blog), 12 August 2019. https://aiimpacts.org/evidence-against-current-methods-leading-to-human-level-artificial-intelligence/; Kirk, Robert, and David Krueger. 'Causal Confusion as an Argument against the Scaling Hypothesis'. AI Alignment Forum, 20 June 2022. https://www.alignmentforum.org/posts/FZL4ftXvcuKmmobmj/causal-confusion-as-an-argument-against-the-scaling. See also Barak, Boaz. 'Injecting Some Numbers into the AGI Debate'. *Windows On Theory* (blog), 27 June 2022. https://windowsontheory.org/2022/06/27/injecting-some-numbers-into-the-agi-debate/. See also: Marcus, Gary. 'What "Game over" for the Latest Paradigm in AI Might Look Like'. Substack newsletter. *The Road to AI We Can Trust* (blog), 29 October 2022. https://garymarcus.substack.com/p/what-game-over-for-the-latest-paradigm (arguing that there will be three limits to "scaling maximalism": insufficient data, insufficient compute, and insufficient task scaling).

[49] Levin, John-Clark, and Matthijs M. Maas. 'Roadmap to a Roadmap: How Could We Tell When AGI Is a "Manhattan Project" Away?', 7. Santiago de Compostela, Spain, 2020. http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_11.pdf.

[50] The taxonomy of approaches presented in this section follows and extends a framework by: Karnofsky, Holden. 'AI Timelines: Where the Arguments, and the "Experts," Stand'. Cold Takes, 7 September 2021. https://www.cold-takes.com/where-ai-forecasting-stands-today/. See also previously Muelhauser, Luke. 'What Do We Know about AI Timelines?' Open Philanthropy. *Open Philanthropy*, 12 October 2015. https://www.openphilanthropy.org/research/what-do-we-know-about-ai-timelines/.

[51] For a recent literature review of work to estimate timelines to advanced, transformative AI systems, see: Wynroe, Keith, David Atkinson, and Jaime Sevilla. 'Literature Review of Transformative Artificial Intelligence Timelines'. Epoch, 17 January 2023. https://epochai.org/blog/literature-review-of-transformative-artificial-intelligence-timelines.

relying more on inside-view arguments,[52] with no specific ranking on the basis of the strength of individual lines of evidence.

Outside-view analyses of timelines

Outside-view analyses of AI development timelines, including:

- → Estimates based on philosophical arguments and anthropic reasoning:
  - → Prima facie likelihood that we (of all generations) are the ones to find ourselves living in the "most important" century, one that we can expect to contain things such as transformative technologies.[53]

- → Estimates based on extrapolating historical (growth) trends:
  - → Insights from endogenous growth theory on AI development dynamics;[54]
  - → Likelihood of explosive economic growth occurring this century, for some reason (plausibly technological, plausibly AI[55]), given analyses of long-run economic history;[56]

[52] The distinction between an "inside view" and an "outside view" analysis used here derives from the classical psychological work by Kahnemann, Tverskey, and Lovallo on planning and forecasting biases. In this model, "[a]n inside view forecast is generated by focusing on the case at hand, by considering the plan and the obstacles to its completion, by constructing scenarios of future progress, and by extrapolating current trends. The outside view [...] essentially ignores the details of the case at hand, and involves no attempt at detailed forecasting of the future history of the project. Instead, it focuses on the statistics of a class of cases chosen to be similar in relevant respects to the present one. The case at hand is also compared to other members of the class, in an attempt to assess its position in the distribution of outcomes for the class." Kahneman, Daniel, and Dan Lovallo. 'Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking'. *Management Science* 39, no. 1 (1993): 17–31. https://econpapers.repec.org/article/inmormnsc/v_3a39_3ay_3a1993_3ai_3a1_3ap_3a17-31.htm. The insight that outside-view forecasts can in some circumstances be more accurate than inside-view evaluations underlies the methodology called "reference class forecasting." In more recent work on "superforecasting" of various events, there is an emphasis on first taking the outside view and only then to modify the conclusion using the inside view. Tetlock, Philip E., and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Reprint edition. Broadway Books, 2016. For an application of these lessons to AI forecasting, see also: Kokotajlo, Daniel. 'Evidence on Good Forecasting Practices from the Good Judgment Project: An Accompanying Blog Post'. AI Impacts, 7 February 2019. https://aiimpacts.org/evidence-on-good-forecasting-practices-from-the-good-judgment-project-an-accompanying-blog-post/. For an argument that integrates inside- and outside-view arguments on AI risk, see: Armstrong, Stuart. 'Is AI an Existential Threat? We Don't Know, and We Should Work on It'. York University, Toronto, 30 November 2020. https://www.youtube.com/watch?v=WLXuZtWoRcE. For a discussion of the pitfalls of unreflexive appeals to "outside view" evaluation, see: Kokotajlo, Daniel. 'Taboo "Outside View"'. EA Forum, 17 June 2021. https://forum.effectivealtruism.org/posts/wYpARcC4WqMsDEmYR/taboo-outside-view.

[53] MacAskill, William. 'Are We Living at the Hinge of History?' Global Priorities Institute, September 2020. https://globalprioritiesinstitute.org/wp-content/uploads/William-MacAskill_Are-we-living-at-the-hinge-of-history.pdf.

[54] Yudkowsky, Eliezer. 'Intelligence Explosion Microeconomics'. Machine Intelligence Research Institute, 2013. http://intelligence.org/files/IEM.pdf.

[55] Davidson, Tom. 'Could Advanced AI Drive Explosive Economic Growth?' Open Philanthropy Project, 8 April 2021. https://www.openphilanthropy.org/could-advanced-ai-drive-explosive-economic-growth.

[56] Roodman, David. 'Modeling the Human Trajectory'. *Open Philanthropy*, 15 June 2020. https://www.openphilanthropy.org/blog/modeling-human-trajectory. See also previously: AI Impacts. 'Precedents for Economic N-Year Doubling before 4n-Year Doubling'. *AI Impacts* (blog), 14 April 2020. https://aiimpacts.org/precedents-for-economic-n-year-doubling-before-4n-year-doubling/. This approach also contains skeptical accounts: see for instance Thorstad, David. 'Against the Singularity Hypothesis'. GPI Working Paper. Global Priorities Institute, University of Oxford, 1 November 2022. https://globalprioritiesinstitute.org/against-the-singularity-hypothesis-david-thorstad/. (arguing that the scenario rests on "implausible growth assumptions").

- → The accelerating historical rate of development of new technologies[57] as well as potential changes in the historical rate of increase in the economy;[58]
- → The historical patterns of barriers to technology development,[59] including unexpected barriers or delays in innovation,[60] as well as lags in subsequent deployment or diffusion.[61]

→ Estimates based on extrapolating from historical trends in efforts dedicated to creating advanced AI:

- → External "semi-informative priors" (i.e., only basic information regarding how long people have attempted to build advanced, transformative AI and what resources they have used, and comparing it to how long it has taken other comparable research fields to achieve their goals given certain levels of funding and effort);[62]
- → Arguments extrapolating from "significantly increased near-future investments in AI progress" given that (comparatively) moderate past investments already yielded significant progress.[63]

→ Estimates based on meta-induction from the track record of past predictions:

[57] Roser, Max. 'Technology over the Long Run: Zoom out to See How Dramatically the World Can Change within a Lifetime'. Our World in Data, 6 December 2022. https://ourworldindata.org/technology-long-run.

[58] Wiblin, Robert, and Keiran Harris. 'Ian Morris on Whether Deep History Says We're Heading for an Intelligence Explosion'. 80,000 Hours Podcast. Accessed 26 October 2023. https://80000hours.org/podcast/episodes/ian-morris-deep-history-intelligence-explosion/.

[59] However, for an older general critique of (naive) attempts to forecast either the boundaries or direction of future technological developments on the basis of historical analogies, see also Stearns, Peter N. 'Forecasting the Future: Historical Analogies and Technological Determinism'. *The Public Historian* 5, no. 3 (1983): 31–54. https://doi.org/10.2307/3377027.

[60] Maas, Matthijs. 'Paths Untaken: The History, Epistemology and Strategy of Technological Restraint, and Lessons for AI'. *Verfassungsblog* (blog), 9 August 2022. https://verfassungsblog.de/paths-untaken/.

[61] Barnett. 'Three Reasons to Expect Long AI Timelines'. LessWrong, 22 April 2021. https://www.lesswrong.com/posts/Z5gPrKTR2oDmm6fqJ/three-reasons-to-expect-long-ai-timelines.; see also Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. https://drive.google.com/file/d/1vIJUAp_i41A5gc9Tb9EvO9aSuLn15ixq/view?usp=sharing. (pg. 76-80).

[62] Davidson, Tom. 'Semi-Informative Priors over AI Timelines'. Open Philanthropy Project, 25 March 2021. https://www.openphilanthropy.org/research/semi-informative-priors-over-ai-timelines/.

[63] Roser, Max. 'Artificial Intelligence Has Advanced despite Having Few Resources Dedicated to Its Development – Now Investments Have Increased Substantially'. Our World in Data, 6 December 2022. https://ourworldindata.org/ai-investments.

→ The general historical track record of past technological predictions, especially those made by futurists[64] as well as those made in professional long-range forecasting exercises,[65] to understand the frequency of over- or underconfidence and of periods of excessive optimism (hype) or excessive pessimism (counterhype);[66]

→ The specific historical track record of past predictions around AI development[67] and the frequency of past periods' excessive optimism (hype) or excessive pessimism (counterhype or "underclaiming"[68]).[69]

[64] On the track record of past futurist predictions of technology: see previously Muelhauser, Luke. 'Futurism's Track Record'. LessWrong, 2014. https://www.lesswrong.com/posts/6ycPKhdmDgWsoyvKd/futurism-s-track-record.; for a more optimistic recent analysis, see: Karnofsky, Holden. 'The Track Record of Futurists Seems ... Fine'. Cold Takes, 30 June 2022. https://www.cold-takes.com/the-track-record-of-futurists-seems-fine/. Karnofsky here draws on an analysis of predictions made by the "Big Three" SF authors (Arthur C Clarke, Robert Heinlein, and Isaac Asimov); Arb Research. 'Scoring the Big 3's Predictive Performance'. Arb, 2022. https://arbresearch.com/files/big_three.pdf. However, for a critical response, see: Luu, Dan. 'Futurist Prediction Methods and Accuracy', 14 September 2022. https://danluu.com/futurist-predictions/ (reviewing the track record of many other influential futurists to argue that the predictive track record of many is quite bad and that more recent longtermist analyses such as Karnofsky's "fundamentally use the same techniques as the futurists analyses we looked at here and then add a few things on top that are also things that people who make accurate predictions do"). For commentary, see also Sempere, Nuño. 'Forecasting Newsletter: September 2022.' Substack newsletter. *Forecasting* (blog), 12 October 2022. https://forecasting.substack.com/p/forecasting-newsletter-september-57b. For a specific discussion of the forecasting track record of some prominent contributors to the AI risk debate, with implications for estimates of AI risk, see also: Garfinkel, Benjamin. 'On Deference and Yudkowsky's AI Risk Estimates'. EA Forum, 19 June 2022. https://forum.effectivealtruism.org/posts/NBgpPaz5vYe3tH4ga/on-deference-and-yudkowsky-s-ai-risk-estimates.

[65] Muelhauser, Luke. 'Evaluation of Some Technology Forecasts from "The Year 2000"'. *Open Philanthropy* (blog), July 2017. https://www.openphilanthropy.org/research/evaluation-of-some-technology-forecasts-from-the-year-2000/. And generally, Muelhauser, Luke. 'How Feasible Is Long-Range Forecasting?' *Open Philanthropy* (blog), 10 October 2019. https://www.openphilanthropy.org/research/how-feasible-is-long-range-forecasting/.

[66] For more discussion of epistemic pitfalls that may steer technology forecasting towards excessive conservatism, see also: Branwen, Gwern. 'Complexity No Bar to AI', 1 June 2014. https://www.gwern.net/Complexity-vs-AI. (Appendix: Technology Forecasting Errors: Functional Fixedness in Assuming Dependencies). See also Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. (pg. 59-64); (reviewing cases of excessive optimism—Human Brain Project, ITER, DARPA's 1983–1993 Strategic Computing Initiative, etc.—as well as cases of sudden, discontinuous progress—Wright Flyer, nuclear fission, nuclear bombs, penicillin—as well as cases where progress occurred that was previously held to be impossible—e.g., Dreyfus' rejection of the very possibility of web search, three years before Google's 2004 IPO. On this basis, he suggests that "it can be useful to consider the epistemic situation of society's relation to various failures of technological prediction. Prima facie, in cases where predictions of a certain new technology repeatedly fail or are postponed, we would expect to see more high-profile and protracted scientific and public debates held over a longer period of time, than in cases where a predicted technology arrives more or less on schedule, or where an unexpected breakthrough occurred (where there may have been little public anticipation in the preceding years). If that is so, we would expect frustrated technological predictions to generally produce a bigger cultural footprint over a longer period of time, than do successful predictions or unexpected breakthroughs. This outsized footprint in turn may shape or skew our idea of technological prediction as being categorically over-optimistic"). (pg. 63-64).

[67] On the track record of past AI predictions: Armstrong, Stuart, Kaj Sotala, and Seán S. Ó hÉigeartaigh. 'The Errors, Insights and Lessons of Famous AI Predictions – and What They Mean for the Future'. *Journal of Experimental & Theoretical Artificial Intelligence* 26, no. 3 (3 July 2014): 317–42. https://doi.org/10.1080/0952813X.2014.895105.; AI Impacts. 'Accuracy of AI Predictions'. *AI Impacts* (blog), 4 June 2015. https://aiimpacts.org/accuracy-of-ai-predictions/.; AI Impacts. 'Similar Predictions'. Accessed 16 August 2022. http://www.aiimpacts.org/ai-timelines/predictions-of-human-level-ai-dates/similar-predictions.

[68] On the concept of "underclaiming" in AI generally, see: Bowman, Samuel R. 'The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail'. arXiv, 10 March 2022. https://doi.org/10.48550/arXiv.2110.08300.

[69] Muelhauser, Luke. 'What Should We Learn from Past AI Forecasts?' *Open Philanthropy* (blog), 1 May 2016. https://www.openphilanthropy.org/research/what-should-we-learn-from-past-ai-forecasts/.

Judgment-based analyses of timelines

Judgment-based analyses of timelines, including:

- → Estimates based on (specialist) expert opinions:
  - → Expert opinion surveys of anticipated rates of progress;[70]
  - → Expert elicitation techniques (e.g., Delphi method).[71]

- → Estimates based on (generalist) estimates from information aggregation mechanisms (financial markets; forecaster prediction markets):[72]
  - → Forecasters' predictions of further AI progress on prediction platforms[73] or forecasting competitions;[74]

[70] For influential surveys of AI experts, see: Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 'When Will AI Exceed Human Performance? Evidence from AI Experts'. *Journal of Artificial Intelligence Research* 62 (2018): 729–54. http://arxiv.org/abs/1705.08807 See also: Michael, Julian, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, et al. 'What Do NLP Researchers Believe? Results of the NLP Community Metasurvey', 2022, 31. https://nlpsurvey.net/nlp-metasurvey-results.pdf ; Zhang, Baobao, Noemi Dreksler, Markus Anderljung, Lauren Kahn, Charlie Giattino, Allan Dafoe, and Michael C. Horowitz. 'Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers'. arXiv, 8 June 2022. https://doi.org/10.48550/arXiv.2206.04132.; Stein-Perlman, Zach, Benjamin Weinstein-Raun, and Katja Grace. '2022 Expert Survey on Progress in AI'. AI Impacts, 4 August 2022. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/. See also: Gruetzemacher, Ross, David Paradice, and Kang Bok Lee. 'Forecasting Extreme Labor Displacement: A Survey of AI Practitioners'. *Technological Forecasting and Social Change*, Forthcoming 2020, 43.; Gruetzemacher, Ross, David Paradice, and Kang Bok Lee. 'Forecasting Transformative AI: An Expert Survey', 16 July 2019. http://arxiv.org/abs/1901.08579.
For older surveys, see also Müller, Vincent C., and Nick Bostrom. 'Future Progress in Artificial Intelligence: A Survey of Expert Opinion'. In *Fundamental Issues of Artificial Intelligence*, 555–72. Springer, 2016. http://www.nickbostrom.com/papers/survey.pdf; Baum, Seth D., Ben Goertzel, and Ted G. Goertzel. 'How Long until Human-Level AI? Results from an Expert Assessment'. *Technological Forecasting and Social Change* 78, no. 1 (January 2011): 185–95. https://doi.org/10.1016/j.techfore.2010.09.006.

[71] Gruetzemacher, Ross, Florian E. Dorner, Niko Bernaola-Alvarez, Charlie Giattino, and David Manheim. 'Forecasting AI Progress: A Research Agenda'. *Technological Forecasting and Social Change* 170 (1 September 2021): 120909. https://doi.org/10.1016/j.techfore.2021.120909.

[72] For a general evaluation of when to expect generalist forecasters in prediction markets to beat domain experts, see: Leech, Gavin, and Mischa Yagudin. 'Comparing Top Forecasters and Domain Experts'. Effective Altruism Forum, 6 March 2022. https://forum.effectivealtruism.org/posts/qZqvBLvR5hX9sEkjR/comparing-top-forecasters-and-domain-experts.

[73] See for instance the questions on the forecasting platform Metaculus: Aguirre, Anthony. 'Will There Be Human-Machine Intelligence Parity Before 2040?' Metaculus, 1 December 2016. https://www.metaculus.com/questions/384/humanmachine-intelligence-parity-by-2040/.; Aguirre, Anthony. 'When Will the First Weakly General AI System Be Devised, Tested, and Publicly Announced?' Metaculus, 18 January 2020. https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/.; Barnett, Matthew. 'When Will the First General AI System Be Devised, Tested, and Publicly Announced?' Metaculus, 23 August 2020. https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/. Though see a discussion of the Metaculus approach: Trazzi, Michaël. 'Alex Lawsen On Forecasting AI Progress'. LessWrong, 6 September 2022. https://www.lesswrong.com/posts/pT86qTHDALskxCXsC/alex-lawsen-on-forecasting-ai-progress.

[74] Steinhardt, Jacob. 'Updates and Lessons from AI Forecasting'. Bounded Regret, 18 August 2021. https://bounded-regret.ghost.io/ai-forecasting/. But for a critique of this approach and how it scores, see: nostalgebraist. 'On "Ai Forecasting: One Year In"'. *Nostalgebraist* (blog), 15 September 2022. https://nostalgebraist.tumblr.com/post/695521414035406848/on-ai-forecasting-one-year-in. For a more general discussion of the challenges of using this approach, see also: Sempere, Nuño. 'Hurdles of Using Forecasting as a Tool for Making Sense of AI Progress'. Measure is Unceasing, 7 November 2023. https://nunosempere.com/blog/2023/11/07/hurdles-forecasting-ai/.

→ Current financial markets' real interest rates, assuming the efficient market hypothesis, suggesting that markets reject short timelines.[75]

Inside-view models on AI timelines

Inside-view models-based analyses of timelines, including:

→ Estimates based on first-principle estimates of minimum resource (compute, investment) requirements for a "transformative" AI system, compared against estimated trends in these resources:
  → The "biological anchors" approach:[76] Comparison with human biological cognition by comparing projected trends in the falling costs of training AI models to the expected minimum amount of computation needed to train an AI model as large as the human brain;[77]
  → The "direct approach":[78] Analysis of empirical neural scaling laws in current AI systems to upper bound the compute needed to train a transformative model. In order to provide estimates of the system's development, this analysis can be combined with estimates of future investment in model training, hardware price-performance, and algorithmic progress[79] as well as with potential barriers in the (future) availability of the data and compute needed to train these models.[80]

[75] Chow, Trevor, Basil Halperin, and J. Zachary Mazlish. 'AGI and the EMH: Markets Are Not Expecting Aligned or Unaligned AI in the next 30 Years'. Effective Altruism Forum, 10 January 2023. https://forum.effectivealtruism.org/posts/8c7LycgtkypkgYjZx/agi-and-the-emh-markets-are-not-expecting-aligned-or.

[76] On the "biological anchors" approach to forecasting, see Cotra, Ajeya. 'Forecasting TAI with Biological Anchors (Draft)'. Open Philanthropy Project, July 2020. https://drive.google.com/drive/folders/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP. For summaries of the report, see: Alexander, Scott. 'Biological Anchors: A Trick That Might Or Might Not Work'. Substack newsletter. *Astral Codex Ten* (blog), 23 February 2022. https://astralcodexten.substack.com/p/biological-anchors-a-trick-that-might. For an overview of other summaries, critiques and responses, see also: Aldred, Will. 'AI Timelines via Bioanchors: The Debate in One Place'. EA Forum, 31 July 2022. https://forum.effectivealtruism.org/posts/NnygBgntvoGSuvsRH/ai-timelines-via-bioanchors-the-debate-in-one-place-1. See also reviews, including: Lin, Jennifer. 'Biological Anchors External Review'. Google Docs, 2022. https://docs.google.com/document/d/1_GqOrCo29qKly1z48-mR86IV7TUDfzaEXxD3lGFQ8Wk/edit?; Hobbhahn, Marius. 'Disagreement with Bio Anchors That Lead to Shorter Timelines'. Effective Altruism Forum, 16 November 2022. https://forum.effectivealtruism.org/posts/gWSa7e2CS7KCu78D8/disagreement-with-bio-anchors-that-lead-to-shorter-timelines.

[77] Cotra's report draws, in part, on Carlsmith, Joseph. 'How Much Computational Power Does It Take to Match the Human Brain?' Open Philanthropy Project, 11 September 2020. https://www.openphilanthropy.org/research/how-much-computational-power-does-it-take-to-match-the-human-brain/. For an older account of human-level hardware, see Grace, Katja. 'Human-Level Hardware Timeline'. *AI Impacts* (blog), 22 December 2017. https://aiimpacts.org/human-level-hardware-timeline/.

[78] Barnett, Matthew, and Tamay Besiroglu. 'Scaling Transformative Autoregressive Models'. Epoch, February 2023. https://epochai.org/files/direct-approach.pdf.; Barnett, Matthew, and Tamay Besiroglu. 'The Direct Approach'. Epoch, 25 April 2023. https://epochai.org/blog/the-direct-approach.

[79] Atkinson, David, Matthew Barnett, Edu Roldán, Ben Cottier, and Tamay Besiroglu. 'Direct Approach Interactive Model'. Epoch, 31 May 2023. https://epochai.org/blog/direct-approach-interactive-model.

[80] Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 'Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning'. arXiv, 25 October 2022. https://doi.org/10.48550/arXiv.2211.04325.; see also dynomight. 'First Principles on AI Progress'. Substack newsletter. DYNOMIGHT INTERNET NEWSLETTER, 6 March 2023. https://dynomight.substack.com/p/scaling?publication_id=327510.

- → Estimates based on direct evaluation of outputs (progress in AI systems' capabilities):
  - → Debates over the significance and implications of specific ongoing AI breakthroughs for further development;[81]
  - → Operationalizing and measuring the generality of existing AI systems.[82]

Methodological debates on AI-timelines analysis

Various methodological debates around AI-timelines analysis:

- → On the potential pitfalls in many of the common methods (forecasting methods,[83] extrapolation, expert predictions[84]) in forecasting AI;
- → On the risk of misinterpreting forecasters who are depending on poor operationalization;[85]
- → On the risk of deference cycles in debates over AI timelines[86] because the opinions and analyses of a small number of people end up tacitly informing the evaluations of a wide range of others in ways that create the impression of many people independently achieving similar conclusions;[87]
- → On the (potentially) limited utility of further discourse over and research into AGI timelines: arguments that all low-hanging fruit may already have been plucked[88] and counterarguments that specific timelines remain relevant to prioritizing strategies.[89]

---

[81] There are too many to discuss in detail; but see informally: Cotra, Ajeya. 'Two-Year Update on My Personal AI Timelines'. AI Alignment Forum, 3 August 2022. https://www.alignmentforum.org/posts/AfH2oPHCApdKicM4m/two-year-update-on-my-personal-ai-timelines.; Alexander, Scott. 'Somewhat Contra Marcus On AI Scaling'. Substack newsletter. Astral Codex Ten (blog), 10 June 2022. https://astralcodexten.substack.com/p/somewhat-contra-marcus-on-ai-scaling. Marcus, Gary. 'Does AI Really Need a Paradigm Shift?' Substack newsletter. The Road to AI We Can Trust (blog), 11 June 2022. https://garymarcus.substack.com/p/does-ai-really-need-a-paradigm-shift.

[82] Burden, John, and Jose Hernandez-Orallo. 'Exploring AI Safety in Degrees: Generality, Capability and Control', In *Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2020)*, 2020, 5. http://ceur-ws.org/Vol-2560/paper21.pdf; Casares, Pablo Antonio Moreno, Bao Sheng Loe, John Burden, Sean hEigeartaigh, and José Hernández-Orallo. 'How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild'. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, no. 5 (28 June 2022): 5295–5303. https://doi.org/10.1609/aaai.v36i5.20466.; and see: Hernández-Orallo, José, Bao Sheng Loe, Lucy Cheke, Fernando Martínez-Plumed, and Seán Ó hÉigeartaigh. 'General Intelligence Disentangled via a Generality Metric for Natural and Artificial Intelligence'. *Scientific Reports* 11, no. 1 (24 November 2021): 22822. https://doi.org/10.1038/s41598-021-01997-7.

[83] On the problem of eliciting expert judgment in conditions where there may not be a relevant reference class of experts to draw from, see generally Morgan, M. Granger. 'Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy'. *Proceedings of the National Academy of Sciences* 111, no. 20 (20 May 2014): 7176–84. https://doi.org/10.1073/pnas.1319946111.

[84] Landau-Taylor, Ben. 'Against AGI Timelines'. *Ben Landau-Taylor* (blog), 12 March 2023. https://benlandautaylor.com/2023/03/12/against-agi-timelines/.

[85] nostalgebraist. 'On "AI Forecasting: One Year In"'. *Nostalgebraist* (blog), 15 September 2022. https://nostalgebraist.tumblr.com/post/695521414035406848/on-ai-forecasting-one-year-in.

[86] Clarke, Sam. 'When Reporting AI Timelines, Be Clear Who You're (Not) Deferring To'. EA Forum, 10 October 2022. https://forum.effectivealtruism.org/posts/FtggfJ2oxNSN8Niix/when-reporting-ai-timelines-be-clear-who-you-re-not.

[87] Clarke, Sam, and mccaffary. 'Deference on AI Timelines: Survey Results'. EA Forum, 31 March 2023. https://forum.effectivealtruism.org/posts/BGFbwca4nfagvB9Xb/deference-on-ai-timelines-survey-results.

[88] Brundage, Miles. 'Why AGI Timeline Research/Discourse Might Be Overrated'. EA Forum, 2022. https://forum.effectivealtruism.org/posts/SEqJoRL5Y8cypFasr/why-agi-timeline-research-discourse-might-be-overrated.

[89] Campos, Simon. 'AGI Timelines in Governance: Different Strategies for Different Timeframes'. EA Forum, 19 December 2022. https://forum.effectivealtruism.org/posts/Pt7MxstXxXHak4wkt/agi-timelines-in-governance-different-strategies-for.

Advanced AI trajectories and early warning signals

A third technical subfield aims at charting the trajectories of advanced AI development, especially the potential for rapid and sudden capability gains, and whether there will be advanced warning signs:

- → Exploring likely AGI "takeoff speeds":[90]
  - → From first principles: arguments in favor of "fast takeoff"[91] vs. arguments for slow(er), more continuous development;[92]
  - → By analogy: exploring historical precedents for sudden disjunctive leaps in technological capabilities.[93]

- → Mapping the epistemic texture of the AI development trajectory in terms of possible advance warning signs of capability breakthroughs[94] or the lack of any such fire alarms.[95]

1.2. Impact models for general social impacts from advanced AI

Various significant societal impacts that could result from advanced AI systems:[96]

- → Potential for advanced AI systems to drive significant, even "explosive" economic growth[97] but also risks of significant inequality or corrosive effects on political discourse;[98]

---

[90] For a discussion of the term, see: Barnett, Matthew. 'Distinguishing Definitions of Takeoff'. AI Alignment Forum, 14 February 2020. https://www.alignmentforum.org/posts/YgNYA6pj2hPSDQiTE/distinguishing-definitions-of-takeoff.
See also Alexander, Scott. 'Yudkowsky Contra Christiano On AI Takeoff Speeds'. Substack newsletter. *Astral Codex Ten* (blog), 4 April 2022. https://astralcodexten.substack.com/p/yudkowsky-contra-christiano-on-ai. For an older account, see: Hanson, Robin, and Eliezer Yudkowsky. 'The Hanson-Yudkowsky AI-Foom Debate', 2008, 741. https://intelligence.org/files/AIFoomDebate.pdf

[91] Muelhauser, Luke. 'Intelligence Explosion FAQ'. Machine Intelligence Research Institute, 2013. https://intelligence.org/ie-faq/. and many others. For a recent argument based on a compute-centric framework, see: Davidson, Tom. 'What a Compute-Centric Framework Says about AI Takeoff Speeds - Draft Report'. EA Forum, 23 January 2023. https://forum.effectivealtruism.org/posts/3vDarp6adLPBTux5g/what-a-compute-centric-framework-says-about-ai-takeoff.

[92] Christiano, Paul. 'Takeoff Speeds'. *The Sideways View* (blog), 24 February 2018. https://sideways-view.com/2018/02/24/takeoff-speeds/. Grace, Katja. 'Likelihood of Discontinuous Progress around the Development of AGI'. *AI Impacts* (blog), 23 February 2018. https://aiimpacts.org/likelihood-of-discontinuous-progress-around-the-development-of-agi/.

[93] AI Impacts. 'Cases of Discontinuous Technological Progress'. *AI Impacts* (blog), 31 December 2014. https://aiimpacts.org/cases-of-discontinuous-technological-progress/. Grace, Katja. 'Discontinuous Progress in History: An Update'. AI Impacts, 13 April 2020. https://aiimpacts.org/discontinuous-progress-in-history-an-update/. See also Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. (pg. 59-64).

[94] Cremer, Carla Zoe, and Jess Whittlestone. 'Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI'. *International Journal of Interactive Multimedia and Artificial Intelligence* 6, no. 5 (2021): 100–109. https://www.ijimai.org/journal/sites/default/files/2021-02/ijimai_6_5_10.pdf

[95] Yudkowsky, Eliezer. 'There's No Fire Alarm for Artificial General Intelligence'. *Machine Intelligence Research Institute* (blog), 14 October 2017. https://intelligence.org/2017/10/13/fire-alarm/. But see also: Grace, Katja. 'Beyond Fire Alarms: Freeing the Groupstruck'. AI Impacts, 26 September 2021. https://aiimpacts.org/beyond-fire-alarms-freeing-the-groupstruck/.

[96] Whittlestone, Jess, and Samuel Clarke. 'AI Challenges for Society and Ethics'. In *The Oxford Handbook of AI Governance*, by Jess Whittlestone and Samuel Clarke, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang. Oxford University Press, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.3.

[97] Erdil, Ege, and Tamay Besiroglu. 'Explosive Growth from AI Automation: A Review of the Arguments'. arXiv, 20 September 2023. https://doi.org/10.48550/arXiv.2309.11690.

[98] Acemoglu, Daron. 'Harms of AI'. Working Paper. Working Paper Series. National Bureau of Economic Research, September 2021. https://doi.org/10.3386/w29247.; Dauvergne, Peter. 'The Globalization of Artificial Intelligence: Consequences for the Politics of Environmentalism'. *Globalizations* 0, no. 0 (30 June 2020): 1–15. https://doi.org/10.1080/14747731.2020.1785670.

→ Significant impacts on scientific progress and innovation;[99]

→ Significant impacts on democracy;[100]

→ Lock-in of harmful socio-political dangers as a result of the increasing role of centralization and optimization;[101]

→ Impacts on geopolitics and international stability.[102]

This is an extensive field that spans a wide range of work, and the above is by no means exhaustive.

1.3. Threat models for extreme risks from advanced AI

A second subcluster of work focuses on understanding the threat models of advanced AI risk,[103] based on indirect arguments for risks, specific threat models for direct catastrophe, or takeover,[104] or on specific threat models for indirect risks.[105]

General arguments for risks from AI

Analyses that aim to explore general arguments (by analogy, on the basis of conceptual argument, or on the basis of empirical evidence from existing AI systems) over whether or why we might have grounds to be concerned about advanced AI.[106]

Analogical arguments for risks

Analogies[107] with historical cases or phenomena in other domains:

---

[99] Clarke, Sam, and Jess Whittlestone. 'A Survey of the Potential Long-Term Impacts of AI: How AI Could Lead to Long-Term Changes in Science, Cooperation, Power, Epistemics and Values'. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 192–202. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3514094.3534131.

[100] Kreps, Sarah, and Doug Kriner. 'How AI Threatens Democracy'. *Journal of Democracy* 34, no. 4 (2023): 122–31. https://muse.jhu.edu/pub/1/article/907693 ; Feldstein, Steven. 'The Consequences of Generative AI for Democracy, Governance and War'. *Survival* 65, no. 5 (3 September 2023): 117–42. https://doi.org/10.1080/00396338.2023.2261260.

[101] Siddarth, Divya, Daron Acemoglu, Danielle Allen, Kate Crawford, James Evans, Michael Jordan, and E. Glen Weyl. 'How AI Fails Us'. Carr Center for Human Rights Policy, December 2021. https://carrcenter.hks.harvard.edu/publications/how-ai-fails-us.

[102] Dafoe, Allan. 'AI Governance: A Research Agenda'. Oxford: Center for the Governance of AI, Future of Humanity Institute, 2018. https://www.fhi.ox.ac.uk/govaiagenda/.

[103] The term "threat model" has been defined by Rohin Shah as: "[a] [c]ombination of a development model that says how we get AGI and a risk model that says how AGI leads to existential catastrophe."; Shah, Rohin. 'The Importance of Threat Models for AI Alignment'. 16 February 2021. https://www.youtube.com/watch?v=VC_J_skJNMs.

[104] The term "takeover" has been defined by Holden Karnofsky as "AI systems disempowering humans entirely, leading to a future that has little to do with anything humans value"; Karnofsky, Holden. 'AI Could Defeat All Of Us Combined'. Cold Takes, 9 June 2022. https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/.; and by Ajeya Cotra as "a possibly violent uprising or coup by AI systems'; Cotra, Ajeya. 'Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover'. AI Alignment Forum, 18 July 2022. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.

[105] For a broad mapping of distinct claims and lines of argument for why AI might pose an extreme or existential risk, see: Hadshar, Rose. 'A Mapping of Claims about AI Risk'. AI Impacts, 18 October 2023. https://blog.aiimpacts.org/p/a-mapping-of-claims-about-ai-risk?publication_id=1465527.

[106] See also Hadshar, Rose. 'A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking'. Research Report. AI Impacts, 2023. https://blog.aiimpacts.org/p/new-report-a-review-of-the-empirical?publication_id=1465527. Note that the distinction in the following sections between conceptual arguments and empirical evidence is drawn from here.

[107] For general work on the role of analogies (whether or not with historical cases) in shaping the agenda and trajectory of AI governance, see also: Maas, Matthijs. 'AI Is Like... A Literature Review of AI Metaphors and Why They Matter for Policy'. AI Foundations Report. Institute for Law & AI, October 2023. https://www.legalpriorities.org/research/ai-policy-metaphors.

- → Historical cases of intelligence enabling control: emergence of human dominion over the natural world: "second species argument"[108] and "the human precedent as indirect evidence of danger";[109]
- → Historical cases where actors were able to achieve large shifts in power despite only wielding relatively minor technological advantages: conquistadors;[110]
- → Historical cases of "lock-in" of suboptimal or bad societal trajectories based on earlier choices and exacerbated by various mechanisms for lock-in: climate change, the agricultural revolution, and colonial projects.[111]

Analogies with known "control problems" observed in other domains:

- → Analogies with economics principal-agent problems;[112]
- → Analogies with constitutional law "incomplete contracting" theorems;[113] in particular, the difficulty of specifying adequate legal responses to all situations or behaviors in advance because it is hard to specify specific and concrete rules for all situations (or in ways that cannot be gamed), whereas vague standards (such as the "reasonable person test") may rely on intuitions that are widely shared but difficult to specify and need to be adjudicated ex post;[114]
- → Analogies to economic systems[115] and to bureaucratic systems and markets, and their accordant failure modes and externalities;[116]
- → Analogies to "Goodhart's Law," where a proxy target metric is used to improve a system so far that further optimization becomes ineffective or harmful;[117]
- → Analogies to the "political control problem"—the problem of the alignment and control of powerful social entities (corporations, militaries, political parties) with (the interests of) their societies, a

[108] Ngo, Richard. 'AGI Safety From First Principles', 2020. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ.
[109] Leech, Gavin. 'Why Worry about Future AI?' Argmin gravitas, 21 March 2021. https://www.gleech.org/ai-risk.
[110] Kokotajlo, Daniel. 'Cortés, Pizarro, and Afonso as Precedents for Takeover'. AI Alignment Forum, 1 March 2020. https://www.alignmentforum.org/posts/ivpKSjM4D6FbqF4pZ/cortes-pizarro-and-afonso-as-precedents-for-takeover.
[111] Clarke, Sam. 'Clarifying "What Failure Looks like" (Part 1)'. AI Alignment Forum, 20 September 2020. https://www.alignmentforum.org/posts/v6Q7T335KCMxujhZu/clarifying-what-failure-looks-like-part-1.
[112] Carlier, Alexis. 'What Can the Principal-Agent Literature Tell Us about AI Risk?' AI Alignment Forum, 8 February 2020. https://www.alignmentforum.org/posts/Z5ZBPEgufmDsm7LAv/what-can-the-principal-agent-literature-tell-us-about-ai.
[113] Hadfield-Menell, Dylan, and Gillian Hadfield. 'Incomplete Contracting and AI Alignment'. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019. http://arxiv.org/abs/1804.04268.
[114] See broadly: Casey, Anthony J., and Anthony Niblett. 'The Death of Rules and Standards'. *Indiana Law Journal* 92, no. 4 (Fall 2017): 1401–47. https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=2444&context=law_and_economics; and for a discussion of how AI systems may also help address or transform this aspect of law, see: Alarie, Benjamin, Anthony Niblett, and Albert H. Yoon. 'Law in the Future'. *University of Toronto Law Journal*, 7 November 2016. https://doi.org/10.3138/UTLJ.4005.; Casey, Anthony J, and Anthony Niblett. 'Self-Driving Laws'. *University of Toronto Law Journal* 66, no. 4 (October 2016): 429–42. https://doi.org/10.3138/UTLJ.4006.
[115] Hadshar, Rose, and particlemania. 'The Economy as an Analogy for Advanced AI Systems'. AI Alignment Forum, 15 November 2022. https://www.alignmentforum.org/posts/oH3XmScSFnZt6x2eN/the-economy-as-an-analogy-for-advanced-ai-systems-2.
[116] Danzig, Richard. 'Machines, Bureaucracies, and Markets as Artificial Intelligences'. Center for Security and Emerging Technology, January 2022. https://cset.georgetown.edu/publication/machines-bureaucracies-and-markets-as-artificial-intelligences/.
[117] Manheim, David, and Scott Garrabrant. 'Categorizing Variants of Goodhart's Law'. *ArXiv:1803.04585 [Cs, q-Fin, Stat]*, 12 March 2018. http://arxiv.org/abs/1803.04585.

problem that remains somewhat unsolved, with societal solutions relying on patchwork and fallible responses that cannot always prevent misalignment (e.g., corporate malfeasance, military coups, or unaccountable political corruption);[118]

→ Analogies with animal behavior, such as cases of animals responding to incentives in ways that demonstrate specification gaming;[119]

→ Illustration with thought experiments and well-established narrative tropes: "sorcerer's apprentice,"[120] "King Midas problem,"[121] and "paperclip maximizer."[122]

Conceptual arguments for risks

Conceptual and theoretical arguments based on existing ML architectures:

→ Arguments based on the workings of modern deep learning systems.[123]

Conceptual and theoretical arguments based on the competitive environment that will shape the evolutionary development of AIs:

→ Arguments suggesting that competitive pressures amongst AI developers may lead the most successful AI agents to likely have (or be given) undesirable traits, which creates risks.[124]

Empirical evidence for risks

Empirical evidence of unsolved alignment failures in existing ML systems, which are expected to persist or scale in more advanced AI systems:[125]

---

[118] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press, 2022. https://docs.google.com/document/d/e/2PACX-1vQOQ0EBIaEu_LaJqWvdPKu8xlmrOCM6h6gq7eFHnN0Y2GPYoodQjLeilxQ8SUwnbVThXc0k_jCIsCX1/pub

[119] Tim G. J. Rduner, and Helen Toner. 'Key Concepts in AI Safety: Specification in Machine Learning'. Center for Security and Emerging Technology, December 2021. https://doi.org/10.51593/20210031. Pg. 3 ("for example, a captive dolphin in Mississippi, upon learning it would be rewarded for bringing trash to its handler, was observed stowing trash in a corner of its habitat and tearing off small pieces to maximize the number of fish it could 'earn.' Like humans and animals, machines respond to the incentives presented to them").

[120] Yudkowsky, Eliezer. 'AI Alignment: Why It's Hard, and Where to Start'. Machine Intelligence Research Institute, 28 December 2016. https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/.

[121] Conn, Ariel. 'Artificial Intelligence and the King Midas Problem'. *Future of Life Institute* (blog), 12 December 2016. https://futureoflife.org/ai/artificial-intelligence-king-midas-problem/.; Russell, Stuart. 'Of Myths and Moonshine'. Reality Club Conversation: The Myth of AI, 2014. https://www.edge.org/conversation/the-myth-of-ai#26015.

[122] Bostrom, Nick. 'Ethical Issues in Advanced Artificial Intelligence'. In *Machine Ethics and Robot Ethics*, by Wendell Wallach and Peter Asaro, 69–75. edited by Wendell Wallach and Peter Asaro, 1st ed. Routledge, 2003. https://doi.org/10.4324/9781003074991-7.

[123] Ngo, Richard. 'The Alignment Problem from a Deep Learning Perspective'. arXiv, 29 August 2022. https://doi.org/10.48550/arXiv.2209.00626. Cotra, Ajeya. 'Why AI Alignment Could Be Hard with Modern Deep Learning'. Cold Takes, 21 September 2021. https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/.; see also the ongoing project: Dewey, Daniel. 'Global Risk from Deep Learning'. Accessed 1 December 2022. https://www.danieldewey.net/risk/index.html.

[124] Hendrycks, Dan. 'Natural Selection Favors AIs over Humans'. arXiv, 28 March 2023. https://doi.org/10.48550/arXiv.2303.16200.

[125] This section draws in part on cases also discussed in: Hadshar, Rose. 'A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking'. Research Report. AI Impacts, 2023. https://blog.aiimpacts.org/p/new-report-a-review-of-the-empirical?publication_id=1465527.

- → "Faulty reward functions in the wild,"[126] "specification gaming,"[127] and reward model overoptimization;[128]
- → "Instrumental convergence,"[129] goal misgeneralization, and "inner misalignment" in reinforcement learning;[130]
- → Language model misalignment[131] and other unsolved safety problems in modern ML,[132] and the harms from increasingly agentic algorithmic systems.[133]

Empirical examples of elements of AI threat models that have already occurred in other domains or with simpler AI systems:

- → Situational awareness: cases where a large language model displays awareness that it is a model, and it can recognize whether it is currently in testing or deployment;[134]
- → Acquisition of a goal to harm society: cases of AI systems being given the outright goal of harming humanity (ChaosGPT);
- → Acquisition of goals to seek power and control: cases where AI systems converge on optimal policies of seeking power over their environment;[135]
- → Self-improvement: examples of cases where AI systems improve AI systems;[136]

---

[126] Amodei, Dario, and Jack Clark. 'Faulty Reward Functions in the Wild'. *OpenAI* (blog), 2016. https://openai.com/blog/faulty-reward-functions/.

[127] Krakovna, Victoria, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zachary Kenton, Jan Leike, and Shane Legg. 'Specification gaming: the flip side of AI ingenuity'. *Deepmind* (blog), 21 April 2020. https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity. See also Krakovna, Victoria. 'Specification Gaming Examples in AI'. *Deep Safety* (blog), 1 April 2018. https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/.

[128] Gao, Leo, John Schulman, and Jacob Hilton. 'Scaling Laws for Reward Model Overoptimization'. arXiv, 19 October 2022. https://doi.org/10.48550/arXiv.2210.10760.

[129] Turner, Alexander Matt, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 'Optimal Policies Tend to Seek Power'. *ArXiv:1912.01683 [Cs]*, 3 December 2021. http://arxiv.org/abs/1912.01683. See also: Turner, Alexander Matt, and Prasad Tadepalli. 'Parametrically Retargetable Decision-Makers Tend To Seek Power'. arXiv, 11 October 2022. https://doi.org/10.48550/arXiv.2206.13477.

[130] Langosco, Lauro, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. 'Goal Misgeneralization in Deep Reinforcement Learning'. arXiv, 7 September 2022. https://doi.org/10.48550/arXiv.2105.14111.; Shah, Rohin, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 'Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals'. arXiv, 2 November 2022. https://doi.org/10.48550/arXiv.2210.01790.

[131] Kenton, Zachary, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 'Alignment of Language Agents'. *ArXiv:2103.14659 [Cs]*, 26 March 2021. http://arxiv.org/abs/2103.14659.

[132] Hendrycks, Dan, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 'Unsolved Problems in ML Safety'. *ArXiv:2109.13916 [Cs]*, 28 September 2021. http://arxiv.org/abs/2109.13916.

[133] Chan, Alan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, et al. 'Harms from Increasingly Agentic Algorithmic Systems'. arXiv, 20 February 2023. https://doi.org/10.48550/arXiv.2302.10329.

[134] Berglund, Lukas, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 'Taken out of Context: On Measuring Situational Awareness in LLMs'. arXiv, 1 September 2023. https://doi.org/10.48550/arXiv.2309.00667. See also Piper, Kelsey. 'Situational Awareness'. Planned Obsolescence, 26 March 2023. https://www.planned-obsolescence.org/situational-awareness/.

[135] Turner, Alexander Matt. 'On Avoiding Power-Seeking by Artificial Intelligence'. arXiv, 23 June 2022. https://doi.org/10.48550/arXiv.2206.11831. Turner, Alexander Matt, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 'Optimal Policies Tend to Seek Power'. *arXiv:1912.01683 [Cs]*, 3 December 2021. http://arxiv.org/abs/1912.01683.

[136] Leech, Gavin. 'Why Worry about Future AI?' Argmin gravitas, 21 March 2021. https://www.gleech.org/ai-risk.

→ Autonomous replication: the ability of simple software to autonomously spread around the internet in spite of countermeasures (various software worms and computer viruses);[137]

→ Anonymous resource acquisition: the demonstrated ability of anonymous actors to accumulate resources online (e.g., Satoshi Nakamoto as an anonymous crypto billionaire);[138]

→ Deception: cases of AI systems deceiving humans to carry out tasks or meet goals.[139]

Direct threat models for direct catastrophe from AI

Work focused at understanding direct existential threat models.[140] This includes:

→ Various overviews and taxonomies of different accounts of AI risk: Barrett & Baum's "model of pathways to risk,"[141] Clarke et al.'s Modelling Transformative AI Risks (MTAIR),[142] Clarke & Martin on "Distinguishing AI Takeover Scenarios,"[143] Clarke & Martin's "Investigating AI Takeover Scenarios,"[144] Clarke's "Classifying Sources of AI X-Risk,"[145] Vold & Harris "How Does Artificial Intelligence Pose an Existential Risk?,"[146] Ngo "Disentangling Arguments for the Importance of AI Safety,"[147] Grace's overview of arguments for existential risk from AI,[148] Nanda's "threat models,"[149] and Kenton et al.;[150]

---

[137] See historically: Kienzle, Darrell M., and Matthew C. Elder. 'Recent Worms: A Survey and Trends'. In *Proceedings of the 2003 ACM Workshop on Rapid Malcode*, 1–10. WORM '03. New York, NY, USA: Association for Computing Machinery, 2003. https://doi.org/10.1145/948187.948189.; Denning, Peter J. 'The Science of Computing: The Internet Worm'. *American Scientist* 77, no. 2 (1989): 126–28. https://www.jstor.org/stable/27855650

[138] Woodside, Thomas. 'Examples of AI Improving AI', 2 October 2023. https://ai-improving-ai.safe.ai/.

[139] OpenAI. 'GPT-4 System Card'. OpenAI, 14 March 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf. See also Park, Peter S., Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 'AI Deception: A Survey of Examples, Risks, and Potential Solutions'. arXiv, 28 August 2023. https://doi.org/10.48550/arXiv.2308.14752.

[140] This taxonomy draws loosely on: Clarke, Sam. 'Classifying Sources of AI X-Risk'. Effective Altruism Forum, 8 August 2022. https://forum.effectivealtruism.org/posts/e55QpEExmtkRjw9CD/classifying-sources-of-ai-x-risk.

[141] Barrett, Anthony M., and Seth D. Baum. 'A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis'. *Journal of Experimental & Theoretical Artificial Intelligence* 29, no. 2 (4 March 2017): 397–414. https://doi.org/10.1080/0952813X.2016.1186228.

[142] Clarke, Sam, Ben Cottier, Aryeh Englander, Daniel Eth, David Manheim, Samuel Dylan Martin, and Issa Rice. 'Modeling Transformative AI Risks (MTAIR) Project -- Summary Report'. arXiv, 19 June 2022. https://doi.org/10.48550/arXiv.2206.09360.

[143] Clarke, Sam, and Samuel Dylan Martin. 'Distinguishing AI Takeover Scenarios'. AI Alignment Forum, 8 September 2021. https://www.alignmentforum.org/posts/qYzqDtoQaZ3eDDyxa/distinguishing-ai-takeover-scenarios.

[144] Martin, Samuel Dylan. 'Investigating AI Takeover Scenarios'. AI Alignment Forum, 17 September 2021. https://www.alignmentforum.org/posts/zkF9PNSyDKusoyLkP/investigating-ai-takeover-scenarios.

[145] Clarke, Sam. 'Classifying Sources of AI X-Risk'. Effective Altruism Forum, 8 August 2022. https://forum.effectivealtruism.org/posts/e55QpEExmtkRjw9CD/classifying-sources-of-ai-x-risk.

[146] Vold, Karina, and Daniel R. Harris. 'How Does Artificial Intelligence Pose an Existential Risk?' In *The Oxford Handbook of Digital Ethics*, 2021. https://doi.org/10.1093/oxfordhb/9780198857815.013.36.

[147] Ngo, Richard. 'Disentangling Arguments for the Importance of AI Safety'. AI Alignment Forum, 2019. https://www.alignmentforum.org/posts/JbcWQCxKWn3y49bNB/disentangling-arguments-for-the-importance-of-ai-safety.

[148] Grace, Katja. 'List of Sources Arguing for Existential Risk from AI'. AI Impacts, 6 August 2022. https://aiimpacts.org/list-of-sources-arguing-for-existential-risk-from-ai/.

[149] Nanda, Neel. 'My Overview of the AI Alignment Landscape: Threat Models'. Alignment Forum, 26 December 2021. https://www.alignmentforum.org/posts/3DFBbPFZyscrAiTKS/my-overview-of-the-ai-alignment-landscape-threat-models.

[150] Kenton, Zachary, Rohin Shah, David Lindner, Vikrant Varma, Victoria Krakovna, Mary Phuong, Ramana Kumar, and Elliot Catt. 'Clarifying AI X-Risk'. Alignment Forum, 1 November 2022. https://www.alignmentforum.org/posts/GctJD5oCDRxCspEaZ/clarifying-ai-x-risk. Summarizing: Kenton, Zachary, Rohin Shah, David Lindner, Vikrant Varma, Victoria Krakovna, Mary Phuong, Ramana Kumar, and Elliot Catt. 'Threat Model Literature Review'. Alignment Forum, 1 November 2022. https://www.alignmentforum.org/posts/wnnkD6P2k2TfHnNmt/threat-model-literature-review.

→ Analysis of potential dangerous capabilities that may be developed by general-purpose AI models, such as cyber-offense, deception, persuasion and manipulation, political strategy, weapons acquisition, long-horizon planning, AI development, situational awareness, and self-proliferation.[151]

Scenarios for direct catastrophe caused by AI

Other lines of work have moved from providing indirect arguments of risk, to instead sketching specific scenarios in and through which advanced AI systems could directly inflict existential catastrophe.

Scenario: Existential disaster because of misaligned superintelligence or power-seeking AI

→ Older accounts, including by Yudkowsky,[152] Bostrom,[153] Sotala,[154] Sotala and Yampolskiy,[155] and Alexander;[156]

→ Newer accounts, such as Cotra & Karnofsky's "AI takeover analysis,"[157] Christiano's account of "What Failure Looks Like,"[158] Carlsmith on existential risks from power-seeking AI,[159] Ngo on "AGI Safety From First Principles,"[160] and "Minimal accounts" of AI takeover scenarios;[161]

→ Skeptical accounts*:* various recent critiques of AI takeover scenarios.[162]

---

[151] Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, et al. 'Model Evaluation for Extreme Risks'. arXiv, 24 May 2023. https://doi.org/10.48550/arXiv.2305.15324.

[152] Yudkowsky, Eliezer. 'Artificial Intelligence as a Positive and Negative Factor in Global Risk.' In *Global Catastrophic Risks*, by Eliezer Yudkowsky, 308–45. New York: Oxford University Press, 2008. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198570509.001.0001/isbn-9780198570509-book-part-21.

[153] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[154] Sotala, Kaj. 'Disjunctive Scenarios of Catastrophic AI Risk'. In *Artificial Intelligence Safety and Security*, edited by Roman V. Yampolskiy, 1st ed., 315–37. First edition. Boca Raton, FL : CRC Press/Taylor & Francis Group, 2018.: Chapman and Hall/CRC, 2018. https://doi.org/10.1201/9781351251389-22.

[155] Sotala, Kaj, and Roman Yampolskiy. 'Risks of the Journey to the Singularity'. In *The Technological Singularity: Managing the Journey*, edited by Victor Callaghan, James Miller, Roman Yampolskiy, and Stuart Armstrong, 11–23. The Frontiers Collection. Berlin, Heidelberg: Springer, 2017. https://doi.org/10.1007/978-3-662-54033-6_2.

[156] Alexander, Scott. 'Superintelligence FAQ', 2016. https://www.lesswrong.com/posts/LTtNXM9shNM9AC2mp/superintelligence-faq.

[157] Cotra, Ajeya. 'Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover'. AI Alignment Forum, 18 July 2022. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.; Karnofsky, Holden. 'AI Could Defeat All Of Us Combined'. Cold Takes, 9 June 2022. https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/.

[158] Christiano, Paul. 'What Failure Looks Like'. *AI Alignment Forum* (blog), 2019. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

[159] Carlsmith, Joseph. 'Is Power-Seeking AI an Existential Risk?' arXiv, April 2021. http://arxiv.org/abs/2206.13353.

[160] Ngo, Richard. 'AGI Safety From First Principles', 2020. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ.

[161] Karnofsky, Holden. 'AI Could Defeat All Of Us Combined'. Cold Takes, 9 June 2022. https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/.; see also: Ricon, Jose Luis. 'Set Sail For Fail? On AI Risk'. *Nintil*, 4 August 2022. https://nintil.com/ai-safety. And see: Clarke, Sam, and Samuel Dylan Martin. 'Distinguishing AI Takeover Scenarios'. AI Alignment Forum, 8 September 2021. https://www.alignmentforum.org/posts/qYzqDtoQaZ3eDDyxa/distinguishing-ai-takeover-scenarios.

[162] Barak, Boaz, and Ben Edelman. 'AI Will Change the World, but Won't Take It over by Playing "3-Dimensional Chess".' *Windows On Theory* (blog), 22 November 2022. https://windowsontheory.org/2022/11/22/ai-will-change-the-world-but-wont-take-it-over-by-playing-3-dimensional-chess/. ; Fodor, James. 'A Critique of AI Takeover Scenarios'. Effective Altruism Forum, 2022. https://forum.effectivealtruism.org/posts/j7X8nQ7YvvA7Pi4BX/a-critique-of-ai-takeover-scenarios

Scenario: Gradual, irretrievable ceding of human power over the future to AI systems

→ Christiano's account of "What Failure Looks Like, (1)."[163]

Scenario: Extreme "suffering risks" because of a misaligned system

→ Various accounts of "worst-case AI safety";[164]

→ Potential for a "suffering explosion" experienced by AI systems.[165]

Scenario: Existential disaster because of conflict between AI systems and multi-system interactions

→ Disasters because of "cooperation failure"[166] or "multipolar failure."[167]

Scenario: Dystopian trajectory lock-in because of misuse of advanced AI to establish and/or maintain totalitarian regimes;

→ Use of advanced AI to establish robust totalitarianism;[168]

→ Use of advanced AI to establish lock-in of the future values.[169]

---

[163] Christiano, Paul. 'What Failure Looks Like'. *AI Alignment Forum* (blog), 2019. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.; discussed in Clarke, Sam. 'Clarifying "What Failure Looks like" (Part 1)'. *AI Alignment Forum* (blog), 2020. https://www.alignmentforum.org/posts/v6Q7T335KCMxujhZu/clarifying-what-failure-looks-like-part-1.

[164] Sotala, Kaj, and Lukas Gloor. 'Superintelligence As a Cause or Cure For Risks of Astronomical Suffering'. *Informatica* 41, no. 4 (27 December 2017). http://www.informatica.si/index.php/informatica/article/view/1877.; Baumann, Tobias. 'An Introduction to Worst-Case AI Safety'. *Reducing Risks of Future Suffering* (blog), 5 July 2018. https://s-risks.org/an-introduction-to-worst-case-ai-safety/.; Baumann, Tobias. 'Focus Areas of Worst-Case AI Safety'. *Reducing Risks of Future Suffering*, 16 September 2017. https://s-risks.org/focus-areas-of-worst-case-ai-safety/; Tomasik, Brian. 'Astronomical Suffering from Slightly Misaligned Artificial Intelligence', 2018. https://reducing-suffering.org/near-miss/.

[165] Metzinger, Thomas. 'Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology'. *Journal of Artificial Intelligence and Consciousness*, 19 February 2021, 1–24. https://doi.org/10.1142/S270507852150003X. Pg. 3.

[166] Clifton, Jesse. 'Cooperation, Conflict, and Transformative Artificial Intelligence - A Research Agenda'. Center on Long-Term Risk, March 2020. https://longtermrisk.org/files/Cooperation-Conflict-and-Transformative-Artificial-Intelligence-A-Research-Agenda.pdf.

[167] Critch, Andrew, and Thomas Krendl Gilbert. 'What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs)'. LessWrong, 1 April 2021. https://www.lesswrong.com/posts/LpM3EAakwYdS6aRKf/what-multipolar-failure-looks-like-and-robust-agent-agnostic.

[168] Winter, Christoph. 'The Challenges of Artificial Judicial Decision-Making for Liberal Democracy'. In *Judicial Decision-Making: Integrating Empirical and Theoretical Perspectives*, edited by Piotr Bystranowski, Bartosz Janik, and Maciej Próchnicki. Springer Nature, 2022. https://papers.ssrn.com/abstract=3933648.; Dafoe, Allan. 'AI Governance: A Research Agenda'. Oxford: Center for the Governance of AI, Future of Humanity Institute, 2018. https://www.fhi.ox.ac.uk/govaiagenda/. Pg 7. On general treatments, see: Caplan, Bryan. 'The Totalitarian Threat'. In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic, 504–19. Oxford University Press, 2008.; Hilton, Benjamin. 'Risks of Stable Totalitarianism'. 80,000 Hours, 23 September 2022. https://80000hours.org/problem-profiles/risks-of-stable-totalitarianism/. For work on related terms such as "digital authoritarianism," see Dragu, Tiberiu, and Yonatan Lupu. 'Digital Authoritarianism and the Future of Human Rights'. *International Organization* 75, no. 4 (ed 2021): 991–1017. https://doi.org/10.1017/S0020818320000624; Wright, Nicholas. 'How Artificial Intelligence Will Reshape the Global Order: The Coming Competition Between Digital Authoritarianism and Liberal Democracy'. *Foreign Affairs*, 10 July 2018. https://www.foreignaffairs.com/articles/world/2018-07-10/how-artificial-intelligence-will-reshape-global-order.; on "AI-tocracy," see: Beraja, Martin, Andrew Kao, David Y Yang, and Noam Yuchtman. 'AI-Tocracy*'. *The Quarterly Journal of Economics*, 13 March 2023, qjad012. https://doi.org/10.1093/qje/qjad012.

[169] Finnveden, Lukas, C. Jess Riedel, and Carl Shulman. 'Artificial General Intelligence and Lock-In', 2022. https://docs.google.com/document/d/1mkLFhxixWdT5peJHq4rfFzq4QbHyfZtANH1nou68q88/edit?.

Scenario: Failures in or misuse of intermediary (non-AGI) AI systems, resulting in catastrophe

→ Deployment of "prepotent" AI systems that are non-general but capable of outperforming human collective efforts on various key dimensions;[170]

→ Militarization of AI enabling mass attacks using swarms of lethal autonomous weapons systems;[171]

→ Military use of AI leading to (intentional or unintentional) nuclear escalation, either because machine learning systems are directly integrated in nuclear command and control systems in ways that result in escalation[172] or because conventional AI-enabled systems (e.g., autonomous ships) are deployed in ways that result in provocation and escalation;[173]

→ Nuclear arsenals serving as an arsenal "overhang" for advanced AI systems;[174]

→ Use of AI to accelerate research into catastrophically dangerous weapons (e.g., bioweapons);[175]

[170] Critch, Andrew, and David Krueger. 'AI Research Considerations for Human Existential Safety (ARCHES)', 29 May 2020. http://acritch.com/arches/. Pg. 12-13 ("We say that an AI system or technology is prepotent [...] (relative to humanity) if its deployment would transform the state of humanity's habitat—currently the Earth—in a manner that is at least as impactful as humanity and unstoppable to humanity, as follows:
➔ at least as impactful as humanity: By this we mean that if the AI system or technology is deployed, then its resulting transformative effects on the world would be at least as significant as humanity's transformation of the Earth thus far, including past events like the agricultural and industrial revolutions.
➔ unstoppable to humanity: By this we mean that if the AI system or technology is deployed, then no concurrently existing collective of humans would have the ability to reverse or stop the transformative impact of the technology (even if every human in the collective were suddenly in unanimous agreement that the transformation should be reversed or stopped). Merely altering the nature of the transformative impact does not count as stopping it.")
[171] Aguirre, Anthony. 'Why Those Who Care about Catastrophic and Existential Risk Should Care about Autonomous Weapons'. EA Forum, 11 November 2020. https://forum.effectivealtruism.org/posts/oR9tLNRSAep293rr5/why-those-who-care-about-catastrophic-and-existential-risk-2.; but for a critical response, see: Ruhl, Christian. 'Risks from Autonomous Weapon Systems and Military AI'. Founders Pledge, 19 May 2022. https://forum.effectivealtruism.org/posts/RKMNZn7r6cT2Yaorf/risks-from-autonomous-weapon-systems-and-military-ai
[172] Maas, Matthijs M, Kayla Matteucci, and Di Cooke. 'Military Artificial Intelligence as Contributor to Global Catastrophic Risk', 2023, in The Era of Global Risk (2023). (eds. SJ Beard, Martin Rees, Catherine Richards & Clarissa Rios-Rojas). Open Book Publishers. 36. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4115010 ; Avin, Shahar, and S. M. Amadae. 'Autonomy and Machine Learning at the Interface of Nuclear Weapons, Computers and People'. In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, edited by V. Boulanin. Stockholm International Peace Research Institute, 2019. https://doi.org/10.17863/CAM.44758.; Rautenbach, Peter. 'Machine Learning & NC3: The Risk of Integration'. Cambridge Existential Risk Initiative, 9 November 2022. https://docs.google.com/document/d/1E2e2gn1LadgwREPb9SfruXq48tPFnd_JNePaXWYSgmo/edit?.
[173] Horowitz, Michael C. 'When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability'. *Journal of Strategic Studies* 42, no. 6 (19 September 2019): 764–88. https://doi.org/10.1080/01402390.2019.1621174. I thank Christian Ruhl for this suggestion.
[174] MichaelA. '8 Possible High-Level Goals for Work on Nuclear Risk'. EA Forum, 29 March 2022. https://forum.effectivealtruism.org/posts/dASEFCurRpNot4Gpc/8-possible-high-level-goals-for-work-on-nuclear-risk.
[175] Clarke, Sam, and Jess Whittlestone. 'A Survey of the Potential Long-Term Impacts of AI: How AI Could Lead to Long-Term Changes in Science, Cooperation, Power, Epistemics and Values'. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 192–202. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3514094.3534131.

→ Use of AI to lower the threshold of access to dual-use biotechnology, creating risks of actors misusing it to create bioweapons.[176]

Other work: vignettes, surveys, methodologies, historiography, critiques

→ Work to sketch vignettes reflecting on potential threat models:
  → AI Impacts' AI Vignettes project;[177]
  → FLI Worldbuilding competition;[178]
  → Wargaming exercises;[179]
  → Other vignettes or risk scenarios.[180]

→ Surveys of how researchers rate the relative probability of different existential risk scenarios from AI;[181]

→ Developing methodologies for AI future developments and risk identification,[182] such as red-teaming,[183] wargaming exercises,[184] and participatory technology assessment,[185] as well as

[176] Soice, Emily H., Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 'Can Large Language Models Democratize Access to Dual-Use Biotechnology?' arXiv, 6 June 2023. https://doi.org/10.48550/arXiv.2306.03809.; Sandbrink, Jonas B. 'Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools'. arXiv, 14 July 2023. https://doi.org/10.48550/arXiv.2306.13952.; Mouton, Christopher A., Caleb Lucas, and Ella Guest. 'The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach'. RAND Corporation, 16 October 2023. https://www.rand.org/pubs/research_reports/RRA2977-1.html.

[177] AI Impacts. 'AI Vignettes Project'. AI Impacts, 12 October 2021. https://aiimpacts.org/ai-vignettes-project/.

[178] ggilgallon. 'FLI Launches Worldbuilding Contest with $100,000 in Prizes'. EA Forum, 17 January 2022. https://forum.effectivealtruism.org/posts/LjExZCPCHnNNTFDfq/fli-launches-worldbuilding-contest-with-usd100-000-in-prizes.; Future of Life Institute. 'About'. *FLI Worldbuilding Contest* (blog), 2022. https://worldbuild.ai/about/.

[179] Avin, Shahar, Ross Gruetzemacher, and James Fox. 'Exploring AI Futures Through Role Play'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8–14. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375817.

[180] See amongst others: Clarke, Sam, and Samuel Dylan Martin. 'Distinguishing AI Takeover Scenarios'. AI Alignment Forum, 8 September 2021. https://www.alignmentforum.org/posts/qYzqDtoQaZ3eDDyxa/distinguishing-ai-takeover-scenarios. Hilton, Benjamin. 'What Could an AI-Caused Existential Catastrophe Actually Look Like?' 80,000 Hours, 15 August 2022. https://80000hours.org/articles/what-could-an-ai-caused-existential-catastrophe-actually-look-like/. Karnofsky, Holden. 'AI Could Defeat All Of Us Combined'. Cold Takes, 9 June 2022. https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/.; see also: Ricon, Jose Luis. 'Set Sail For Fail? On AI Risk'. *Nintil*, 4 August 2022. https://nintil.com/ai-safety. Branwen, Gwern. 'It Looks Like You're Trying To Take Over The World', 6 March 2022. https://www.gwern.net/fiction/Clippy. Nielsen, Michael. 'Notes on Existential Risk from Artificial Superintelligence', 18 September 2023. https://michaelnotebook.com/xrisk/index.html.

[181] Carlier, Alexis, Sam Clarke, and Jonas Schuett. 'Existential Risks from AI: A Survey of Expert Opinion', 2021, 16.; Clarke, Sam, Alexis Carlier, and Jonas Schuett. 'Survey on AI Existential Risk Scenarios'. Effective Altruism Forum, 8 June 2021. https://forum.effectivealtruism.org/posts/2tumunFmjBuXdfF2F/survey-on-ai-existential-risk-scenarios-1. Bensinger, Rob. '"Existential Risk from AI" Survey Results'. AI Alignment Forum, 1 June 2021. https://www.alignmentforum.org/posts/QvwSr5LsxyDeaPK5s/existential-risk-from-ai-survey-results. and see indirectly: Graham, Ross. 'Discourse Analysis of Academic Debate of Ethics for AGI'. *AI & SOCIETY*, 2 June 2021. https://doi.org/10.1007/s00146-021-01228-7.

[182] Shahar, Avin. 'Exploring Artificial Intelligence Futures'. *Journal of AI Humanities* 2 (31 October 2018): 169–94. https://doi.org/10.46397/JAIH.2.7.

[183] Hicks, Marie-Laure, Ella Guest, Jess Whittlestone, Jacob Ohrvik-Stott, Sana Zakaria, Cecilia Ang, Chryssa Politi, Imogen Wade, and Salil Gunashekar. 'Exploring Red Teaming to Identify New and Emerging Risks from AI Foundation Models'. RAND Corporation, 31 October 2023. https://www.rand.org/pubs/conf_proceedings/CFA3031-1.html.

[184] Avin, Shahar, Ross Gruetzemacher, and James Fox. 'Exploring AI Futures Through Role Play'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8–14. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375817.

[185] Cremer, Carla Zoe, and Jess Whittlestone. 'Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI'. *International Journal of Interactive Multimedia and Artificial Intelligence* 6, no. 5 (2021): 100–109. https://www.ijimai.org/journal/sites/default/files/2021-02/ijimai_6_5_10.pdf

established risk identification techniques (scenario analysis, fishbone method, and risk typologies and taxonomies), risk analysis techniques (causal mapping, Delphi technique, cross-impact analysis, bow tie analysis, and system-theoretic process analysis), and risk evaluation techniques (checklists and risk matrices);[186]

- → Historiographic accounts of changes in AI risk arguments and debates over time:
  - → General history of concerns around AI risk (1950s–present);[187]
  - → Early history of the rationalist and AI risk communities (1990s–2010);[188]
  - → Recent shifts in arguments (e.g., 2014–present);[189]
  - → Development and emergence of AI risk "epistemic community."[190]

- → Critical investigations of and counterarguments to the case for extreme AI risks, including object-level critiques of the arguments for risk[191] as well as epistemic arguments, arguments about community dynamics, and argument selection effects.[192]

[186] Koessler, Leonie, and Jonas Schuett. 'Risk Assessment at AGI Companies: A Review of Popular Risk Assessment Techniques from Other Safety-Critical Industries'. arXiv, 17 July 2023. http://arxiv.org/abs/2307.08823. For an application of fault trees and influence diagrams to risk analysis, see also Barrett, Anthony M., and Seth D. Baum. 'A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis'. *Journal of Experimental & Theoretical Artificial Intelligence* 29, no. 2 (4 March 2017): 397–414. https://doi.org/10.1080/0952813X.2016.1186228.

[187] Burden, John, Sam Clarke, and Jess Whittlestone. 'From Turing's Speculations to an Academic Discipline: A History of AI Existential Safety'. In *Cambridge Conference on Catastrophic Risk 2020*, 2022; see also: lukeprog. 'AI Risk and Opportunity: Humanity's Efforts So Far'. Accessed 2 February 2023. https://www.lesswrong.com/posts/i4susk4W3ieR5K92u/ai-risk-and-opportunity-humanity-s-efforts-so-far.; Muelhauser, Luke. 'AI Risk & Opportunity: A Timeline of Early Ideas and Arguments'. Accessed 2 February 2023. https://www.lesswrong.com/posts/Qdq2SKyMi8vf7Snxq/ai-risk-and-opportunity-a-timeline-of-early-ideas-and.

[188] Chivers, Tom. *The AI Does Not Hate You: Superintelligence, Rationality and the Race to Save the World*. London: Weidenfeld & Nicolson, 2019.

[189] Adamczewski, Tom. 'A Shift in Arguments for AI Risk'. Fragile Credences, 25 May 2019. https://fragile-credences.github.io/prioritising-ai/.

[190] Ahmed, Shazeda, Klaudia Jazwinska, Archana Ahlawat, Amy Winecoff, and Mona Wang. 'Building the Epistemic Community of AI Safety', 2023. https://drive.google.com/file/d/1HIwKMnQNYme2U4__T-5MvKh9RZ7-RD6x/view?usp=drivesdk.

[191] Garfinkel, Ben. 'How Sure Are We about This AI Stuff?' EA Forum, 9 February 2019. https://forum.effectivealtruism.org/posts/9sBAW3qKppnoG3QPq/ben-garfinkel-how-sure-are-we-about-this-ai-stuff. Grace, Katja. 'Counterarguments to the Basic AI X-Risk Case'. AI Impacts, 14 October 2022. https://aiimpacts.org/counterarguments-to-the-basic-ai-x-risk-case/.; and response: Jenner, Erik, and Johannes Treutlein. 'Response to Katja Grace's AI x-Risk Counterarguments'. Alignment Forum, 19 October 2022. https://www.alignmentforum.org/posts/GQat3Nrd9CStHyGaq/response-to-katja-grace-s-ai-x-risk-counterarguments. Barak, Boaz, and Ben Edelman. 'AI Will Change the World, but Won't Take It over by Playing "3-Dimensional Chess".' *Windows On Theory* (blog), 22 November 2022. https://windowsontheory.org/2022/11/22/ai-will-change-the-world-but-wont-take-it-over-by-playing-3-dimensional-chess/.

[192] See e.g. Heninger, Jeffrey. 'Against a General Factor of Doom'. AI Impacts, 23 November 2022. https://aiimpacts.org/against-a-general-factor-of-doom/. Trammell, Philip. 'But Have They Engaged with the Arguments?' Philip Trammell, 29 December 2019. https://philiptrammell.com/blog/46; NunoSempere. 'My Highly Personal Skepticism Braindump on Existential Risk from Artificial Intelligence.' EA Forum, 23 January 2023. https://forum.effectivealtruism.org/posts/L6ZmggEJw8ri4KB8X/my-highly-personal-skepticism-braindump-on-existential-risk.

Threat models for indirect AI contributions to existential risk factors

Work focused at understanding indirect ways in which AI could contribute to existential threats, such as by shaping societal "turbulence"[193] and other existential risk factors.[194] This covers various long-term impacts on societal parameters such as science, cooperation, power, epistemics, and values:[195]

- → Destabilizing political impacts from AI systems in areas such as domestic politics (e.g., polarization, legitimacy of elections), international political economy, or international security[196] in terms of the balance of power, technology races and international stability, and the speed and character of war ;
- → Hazardous malicious uses;[197]
- → Impacts on "epistemic security" and the information environment;[198]
- → Erosion of international law and global governance architectures;[199]
- → Other diffuse societal harms.[200]

1.4. Profile of technical alignment problem

- → Work mapping different geographical or institutional hubs active on AI alignment: overview of the AI safety community and problem,[201] and databases of active research institutions[202] and of research;[203]

[193] Bostrom, Nick, Allan Dafoe, and Carrick Flynn. 'Public Policy and Superintelligent AI: A Vector Field Approach'. In *Ethics of Artificial Intelligence*, edited by S.M. Liao. Oxford University Press, 2019. http://www.nickbostrom.com/papers/aipolicy.pdf.

[194] Ord, Toby. *The Precipice: Existential Risk and the Future of Humanity*. Illustrated Edition. New York: Hachette Books, 2020. pg. 175–180.

[195] Clarke, Sam, and Jess Whittlestone. 'A Survey of the Potential Long-Term Impacts of AI: How AI Could Lead to Long-Term Changes in Science, Cooperation, Power, Epistemics and Values'. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 192–202. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3514094.3534131.

[196] For an overview of some of these themes, see also: Dafoe, Allan. 'AI Governance: Opportunity and Theory of Impact', 17 September 2020. https://www.allandafoe.com/opportunity. *'AI Strategy, Policy, and Governance' by Allan Dafoe*, 2019. https://www.youtube.com/watch?v=2IpJ8TIKKtI.

[197] Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', 20 February 2018. http://arxiv.org/abs/1802.07228.

[198] Seger, Elizabeth, Shahar Avin, Gavin Pearson, Mark Briers, Seán Ó hÉigeartaigh, and Helena Bacon. 'Tackling Threats to Informed Decisionmaking in Democratic Societies: Promoting Epistemic Security in a Technologically-Advanced World'. The Alan Turing Institute, October 2020. https://www.turing.ac.uk/research/publications/tackling-threats-informed-decision-making-democratic-societies.

[199] Maas, Matthijs M. 'International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order'. *Melbourne Journal of International Law* 20, no. 1 (2019): 29–56. https://law.unimelb.edu.au/__data/assets/pdf_file/0005/3144308/Maas.pdf

[200] Kolt, Noam. 'Algorithmic Black Swans'. *Washington University Law Review* 101 (25 February 2023). https://papers.ssrn.com/abstract=4370566.

[201] Hilton, Benjamin. 'Preventing an AI-Related Catastrophe - Problem Profile'. 80,000 Hours, 25 August 2022. https://80000hours.org/problem-profiles/artificial-intelligence/.

[202] Aird, Michael. 'Database of Orgs Relevant to Longtermist/x-Risk Work'. EA Forum, 19 November 2021. https://forum.effectivealtruism.org/posts/twMs8xsgwnYvaowWX/database-of-orgs-relevant-to-longtermist-x-risk-work. (see link to database).

[203] Riedel, Jess, and Angelica Deibel. 'TAI Safety Bibliographic Database'. AI Alignment Forum, 22 December 2020. https://www.alignmentforum.org/posts/4DegbDJJiMX2b3EKm/tai-safety-bibliographic-database.

→ Work mapping current technical alignment approaches;[204]

→ Work aiming to assess the (relative) efficacy or promise of different approaches to alignment, insofar as possible:[205] Cotra,[206] Soares,[207] and Leike.[208]

→ Mapping the relative contributions to technical AI safety by different communities[209] and the chance that AI safety problems get "solved by default";[210]

→ Work mapping other features of AI safety research, such as the need for minimally sufficient access to AI models under API-based "structured access" arrangements.[211]

2. Deployment parameters

Another major part of the field aims to understand the parameters of the advanced AI deployment landscape by mapping the size and configuration of the "game board" of relevant advanced AI developers—the actors whose (ability to take) key decisions (e.g., around whether or how to deploy particular advanced AI systems,

[204] Ji, Jiaming, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, et al. 'AI Alignment: A Comprehensive Survey'. arXiv, 1 November 2023. https://doi.org/10.48550/arXiv.2310.19852. See more specifically Leech, Gavin. 'A largely uncategorised list of live alignment agendas', 2023, https://arbresearch.com/files/agendas_2023.pdf. Any review of work in this space will be incomplete and rapidly out of date. However, for a sample of slightly older work, see: Nanda, Neel. 'My Overview of the AI Alignment Landscape: Threat Models'. Alignment Forum, 26 December 2021. https://www.alignmentforum.org/posts/3DFBbPFZyscrAiTKS/my-overview-of-the-ai-alignment-landscape-threat-models. Krakovna, Victoria. 'Paradigms of AI Alignment: Components and Enablers'. *Victoria Krakovna* (blog), 2 June 2022. https://vkrakovna.wordpress.com/2022/06/02/paradigms-of-ai-alignment-components-and-enablers/. Kirchner, Jan Hendrik, Logan Riggs Smith, Jacques Thibodeau, and janus. 'A Descriptive, Not Prescriptive, Overview of Current AI Alignment Research'. AI Alignment Forum, 6 June 2022. https://www.alignmentforum.org/posts/FgjcHiWvADgsocE34/a-descriptive-not-prescriptive-overview-of-current-ai. Hubinger, Evan. 'An Overview of 11 Proposals for Building Safe Advanced AI'. arXiv, 4 December 2020. https://doi.org/10.48550/arXiv.2012.07532. Christiano, Paul. 'Current Work in AI Alignment'. Effective Altruism, 3 April 2020. https://www.effectivealtruism.org/articles/paul-christiano-current-work-in-ai-alignment. Everitt, Tom, Gary Lea, and Marcus Hutter. 'AGI Safety Literature Review'. *ArXiv:1805.01109 [Cs]*, 3 May 2018. http://arxiv.org/abs/1805.01109.

[205] As in many fields at an early stage of development, there may be significant challenges to meaningfully evaluating or comparing the relative promises of different paradigms of alignment research. As such, while some assessment of past work can compare the evaluations of different approaches, any larger comparisons of these agendas will be quite precarious. I thank Richard Ngo for this point.

[206] Cotra, Ajeya. 'Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover'. AI Alignment Forum, 18 July 2022. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to. (see subsection 'Simple "baseline" behavioral safety interventions').

[207] Soares, Nate. 'On How Various Plans Miss the Hard Bits of the Alignment Challenge'. Effective Altruism Forum, 12 July 2022. https://forum.effectivealtruism.org/posts/jydymb23NWF3Q4oDt/on-how-various-plans-miss-the-hard-bits-of-the-alignment.

[208] Leike, Jan. 'Why I'm Optimistic about Our Alignment Approach'. *Musings on the Alignment Problem* (blog), 5 December 2022. https://aligned.substack.com/p/alignment-optimism?publication_id=328633&isFreemail=true.

[209] Leech, Gavin. 'The Academic Contribution to AI Safety Seems Large'. Effective Altruism Forum, 2020. https://forum.effectivealtruism.org/posts/8ErtxW7FRPGMtDqJy/the-academic-contribution-to-ai-safety-seems-large.

[210] Shah, Rohin. '[AN #80]: Why AI Risk Might Be Solved without Additional Intervention from Longtermists'. AI Alignment Forum, 2 January 2020. https://www.alignmentforum.org/posts/QknPz9JQTQpGdaWDp/an-80-why-ai-risk-might-be-solved-without-additional.

[211] Bucknall, Benjamin S, and Robert F Trager. 'Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements'. Oxford Martin AI Governance Initiative, October 2023. https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements/.

how much to invest in alignment research, etc.) may be key in determining risks and outcomes from advanced AI.

As such, there is significant work on mapping the disposition of the AI development ecosystem and how this will determine who is (or will likely be) in the position to develop and deploy the most advanced AI systems. Some work in this space focuses on mapping the current state of these deployment parameters; other work focuses on the likely future trajectories of these deployment parameters over time.

2.1. Size, productivity, and geographic distribution of the AI research field

→ Mapping the current size, activity, and productivity of the AI research field;[212]

→ Mapping the global geographic distribution of active AGI programs,[213] including across key players such as the US or China.[214]

2.2. Geographic distribution of key inputs in AI development

→ Mapping the current distribution of relevant inputs in AI development, such as the distribution of computation,[215] semiconductor manufacturing,[216] AI talent,[217] open-source machine learning software,[218] etc.

[212] For an older (2014) estimate, see: Muehlhauser, Luke. 'How Big Is the Field of Artificial Intelligence? (Initial Findings)'. Machine Intelligence Research Institute, 28 January 2014. https://intelligence.org/2014/01/28/how-big-is-ai/.
[213] Fitzgerald, McKenna, Aaron Boddy, and Seth D. Baum. '2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy'. Global Catastrophic Risk Institute Technical Report. Global Catastrophic Risk Institute, 2020. https://gcrinstitute.org/papers/055_agi-2020.pdf. And for a previous 2017 version: Baum, Seth. 'A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy'. Global Catastrophic Risk Institute Technical Report. Global Catastrophic Risk Institute, 2017. https://papers.ssrn.com/abstract=3070741.
[214] See for instance: Hannas, William, Huey-Meei Chang, Daniel Chou, and Brian Fleeger. 'China's Advanced AI Research: Monitoring China's Paths to "General" Artificial Intelligence'. Center for Security and Emerging Technology, July 2022. https://cset.georgetown.edu/publication/chinas-advanced-ai-research/. And also: Hannas, William, Huey-Meei Chang, Catherine Aiken, and Daniel Chou. 'China AI-Brain Research: Brain-Inspired AI, Connectomics, Brain-Computer Interfaces'. Center for Security and Emerging Technology, September 2020. https://cset.georgetown.edu/publication/china-ai-brain-research/. (focusing on connectionist approaches).
[215] For an older, very outdated sketch, see: Muehlhauser, Luke. 'The World's Distribution of Computation (Initial Findings)'. Machine Intelligence Research Institute, 1 March 2014. https://intelligence.org/2014/02/28/the-worlds-distribution-of-computation-initial-findings/. For a project guide to a continued project, see: Grace, Katja, and Luke Muelhauser. 'Project Guide: Map the Computing Landscape'. Google Docs. Accessed 28 October 2022. https://docs.google.com/document/d/19K37J6VzN7aigZC4IwydEWDAYMSVBTWFcrxN6YFMxig/edit?usp=sharing&.
[216] Khan, Saif. 'The Semiconductor Supply Chain: Assessing National Competitiveness'. Center for Security and Emerging Technology, January 2021. https://cset.georgetown.edu/publication/the-semiconductor-supply-chain/.
[217] Huang, Tina, and Zachary Arnold. 'Immigration Policy and the Global Competition for AI Talent'. Center for Security and Emerging Technology, June 2020. https://cset.georgetown.edu/research/immigration-policy-and-the-global-competition-for-ai-talent/.
[218] Langenkamp, Max, and Daniel N. Yue. 'How Open Source Machine Learning Software Shapes AI'. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 385–95. Oxford United Kingdom: ACM, 2022. https://doi.org/10.1145/3514094.3534167.

- → Mapping and forecasting trends in relevant inputs for AI,[219] such as:
  - → Trends in compute inputs scaling[220] and in the training costs and GPU price-performance of machine learning systems over time;[221]
  - → Trends in dataset scaling and potential ceilings;[222]
  - → Trends in algorithmic progress, including their effect on the ability to leverage other inputs, e.g., the relative importance of CPUs versus specialized hardware;[223]
- → Mapping and forecasting trends in input criticality for AI, such as trends in data efficiency[224] and the degree to which data becomes the operative constraint on language model performance.[225]

[219] Epoch. 'Announcing Epoch: A Research Initiative Investigating the Road to Transformative AI', 27 June 2022. https://epochai.org/blog/announcing-epoch.

[220] Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 'Compute Trends Across Three Eras of Machine Learning'. *ArXiv:2202.05924 [Cs]*, 11 February 2022. http://arxiv.org/abs/2202.05924. See also Lohn, Andrew, and Micah Musser. 'AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?' Center for Security and Emerging Technology, January 2022. https://cset.georgetown.edu/publication/ai-and-compute/.

[221] Cottier, Ben. 'Trends in the Dollar Training Cost of Machine Learning Systems'. Epoch, 31 January 2023. https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems.

[222] Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 'Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning'. arXiv, 25 October 2022. https://doi.org/10.48550/arXiv.2211.04325.

[223] Kirchner, Jan Hendrik. 'Compute Governance: The Role of Commodity Hardware'. *On Brains, Minds, And Their Possible Uses* (blog), 26 March 2022. https://universalprior.substack.com/p/compute-governance-the-role-of-commodity.

[224] Tucker, Aaron D., Markus Anderljung, and Allan Dafoe. 'Social and Governance Implications of Improved Data Efficiency'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 378–84. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375863.

[225] Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, et al. 'Training Compute-Optimal Large Language Models'. arXiv, 29 March 2022. http://arxiv.org/abs/2203.15556.

2.3. Organization of global AI supply chain

→ Mapping the current shape of the AI supply chain;[226]

→ Mapping and forecasting dominant actors in the future AI ecosystem, in terms of:
  → different actors' control of and access to key inputs and/or chokepoints;[227]
  → future shape of the AI supply chain (e.g., level of integration and monopoly structure);[228]
  → shape of AI deployment landscape (e.g., dominance of key operators of generative models vs. copycat models).

2.4. Dispositions and values of advanced AI developers

→ Anticipating the likely behavior or attitude of key advanced AI actors with regard to their caution about and investment in safety research, such as expecting AI companies to "race forward" and dedicate "naive safety effort."[229]

2.5. Developments in converging technologies

→ Mapping converging developments in adjacent, potentially intersecting or relevant technologies, such as cryptography,[230] nanotechnology,[231] and others.

---

[226] Küspert, Sabrina, Nicolas Moës, and Connor Dunlop. 'The Value Chain of General-Purpose AI'. Ada Lovelace Institute, 10 February 2023. https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/. Engler, Alex. 'A Typology of the Machine Learning Value Chain — And Why It Matters to Policymaking'. *Brookings* (blog), 20 September 2022. https://www.brookings.edu/blog/techtank/2022/09/20/a-typology-of-the-machine-learning-value-chain-and-why-it-matters-to-policymaking/. Cobbe, Jennifer, Michael Veale, and Jatinder Singh. 'Understanding Accountability in Algorithmic Supply Chains'. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1186–97, 2023. https://doi.org/10.1145/3593013.3594073.

[227] Barbe, Andre, and Will Hunt. 'Preserving the Chokepoints: Reducing the Risks of Offshoring Among U.S. Semiconductor Manufacturing Equipment Firms'. Center for Security and Emerging Technology, May 2022. https://cset.georgetown.edu/publication/preserving-the-chokepoints/.; Murphy, Ben. 'Chokepoints: China's Self-Identified Strategic Technology Import Dependencies'. Center for Security and Emerging Technology, May 2022. https://cset.georgetown.edu/publication/chokepoints/.

[228] Salisbury, Adam. 'How Will the AI Supply Chain Evolve?', 2022. https://docs.google.com/document/d/1s3QGFJ8Ochosksl4JgQCWekJrsY3YFAfGgEiEt6zFpA/edit?usp=sharing&. (arguing the AI supply chain is currently mostly vertically integrated, with the main users of AI technology also producing the majority of their AI capabilities in-house, but reviewing several trends to anticipate "the emergence of a hybrid industry structure in which i) AI firms sell access to some of their technology ii) some non-AI firms develop their own AI capabilities and iii) AI firms retain a major downstream presence themselves."). See also: Mindermann, Sören. 'Summary: Will AI Companies Sell or Use Their Technology? V2'. Google Docs, 2021. https://docs.google.com/document/d/1ltCvbI8xYQ49izUUjbHLlQI6MqS6wCJFsIb6zCjSKMA/edit?. (drawing on established economic theory and historical evidence from general-purpose technologies (GPTs) to argue that the AI industry will likely become less vertically integrated); see also: Uuk, Risto. 'Emerging Non-European Monopolies in the Global AI Market'. Future of Life Institute, November 2022. https://futureoflife.org/wp-content/uploads/2022/11/Emerging_Non-European_Monopolies_in_the_Global_AI_Market.pdf.

[229] Cotra, Ajeya. 'Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover'. AI Alignment Forum, 18 July 2022. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to. See also: Karnofsky, Holden. 'How Might We Align Transformative AI If It's Developed Very Soon?' EA Forum, 29 August 2022. https://forum.effectivealtruism.org/posts/sW6RggfddDrcmM6Aw/how-might-we-align-transformative-ai-if-it-s-developed-very.

[230] Garfinkel, Benjamin. 'A Tour of Emerging Cryptographic Technologies: What They Are and How They Could Matter'. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, May 2021. https://assets.website-files.com/614b70a71b9f71c9c240c7a7/617938781d1308004d007e2d_Garfinkel_Tour_Of_Emerging_Cryptographic_Technologies.pdf.

[231] Snodin, Ben. 'My Thoughts on Nanotechnology Strategy Research as an EA Cause Area'. EA Forum, 2 May 2022. https://forum.effectivealtruism.org/posts/oqBJk2Ae3RBegtFfn/my-thoughts-on-nanotechnology-strategy-research-as-an-ea.

3. Governance parameters

Work on governance parameters aims to map (1) how AI systems are currently being governed, (2) how they are likely to be governed by default (given prevailing perceptions and regulatory initiatives), as well as (3) the conditions for developing and implementing productive governance interventions on advanced AI risk.

Some work in this space focuses on mapping the current state of these governance parameters and how they affect AI governance efforts initiated today. Other work focuses on the likely future trajectories of these governance parameters.

3.1. Stakeholder perceptions of AI

Surveys of current perceptions of AI among different relevant actors:

- → Public perceptions of the future of AI,[232] of AI's societal impacts,[233] of the need for caution and/or regulation of AI,[234] and of the rights or standing of AI entities;[235]
- → Policymaker perceptions of AI[236] and the prominence of different memes, rhetorical frames, or narratives around AI;[237]

[232] Zhang, Baobao, and Allan Dafoe. 'Artificial Intelligence: American Attitudes and Trends'. Center for the Governance of AI and Future of Humanity Institute, January 2019. https://www.ssrn.com/abstract=3312874.

[233] Zhang, Baobao. 'No Rage Against the Machines: Threat of Automation Does Not Change Policy Preferences'. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 856–66. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3514094.3534179.; see also Cave, Stephen, Kate Coughlan, and Kanta Dihal. '"Scary Robots": Examining Public Responses to AI'. In *Proceedings of AAAI / ACM Conference on Artificial Intelligence, Ethics and Society 2019*, 8, 2019. https://dl.acm.org/doi/abs/10.1145/3306618.3314232 ; O'Shaughnessy, Matthew, Daniel S. Schiff, Lav R. Varshney, Christopher Rozell, and Mark Davenport. 'What Governs Attitudes toward Artificial Intelligence Adoption and Governance?' OSF Preprints, 14 December 2021. https://doi.org/10.31219/osf.io/pkeb8.

[234] Dreksler, Noemi, David McCaffary, Lauren Kahn, Kate Mays, Markus Anderljung, Allan Dafoe, Michael C. Horowitz, and Baobao Zhang. 'Preliminary Survey Results: US and European Publics Overwhelmingly and Increasingly Agree That AI Needs to Be Managed Carefully'. GovAI Blog, 17 April 2023. https://www.governance.ai/post/increasing-consensus-ai-requires-careful-management.; O'Shaughnessy, Matthew, Daniel S. Schiff, Lav R. Varshney, Christopher Rozell, and Mark Davenport. 'What Governs Attitudes toward Artificial Intelligence Adoption and Governance?' OSF Preprints, 14 December 2021. https://doi.org/10.31219/osf.io/pkeb8.; Stein-Perlman, Zach. 'The Public Supports Regulating AI for Safety'. AI Impacts, 16 February 2023. https://aiimpacts.org/the-public-supports-regulating-ai-for-safety/. Citing: Monmouth University Poll. 'National: Artificial Intelligence Use Prompts Concerns'. Monmouth University, 15 February 2023. https://www.monmouth.edu/polling-institute/documents/monmouthpoll_us_021523.pdf/.

[235] Martínez, Eric, and Christoph Winter. 'Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection'. *Frontiers in Robotics and AI* 8 (2021). https://www.frontiersin.org/articles/10.3389/frobt.2021.788355.; see also generally: De Graaf, Maartje M. A., Frank A. Hindriks, and Koen V. Hindriks. 'Who Wants to Grant Robots Rights?' *Frontiers in Robotics and AI* 8 (2022). https://www.frontiersin.org/articles/10.3389/frobt.2021.781985.

[236] Krafft, P. M., Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 'Defining AI in Policy versus Practice'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 72–78. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375835.

[237] Imbrie, Andrew, James Dunham, Rebecca Gelles, and Catherine Aiken. 'Mainframes: A Provisional Analysis of Rhetorical Frames in AI'. Center for Security and Emerging Technology, August 2020. https://cset.georgetown.edu/research/mainframes-a-provisional-analysis-of-rhetorical-frames-in-ai/. ; Imbrie, Andrew, Rebecca Gelles, James Dunham, and Catherine Aiken. 'Contending Frames: Evaluating Rhetorical Dynamics in AI'. Center for Security and Emerging Technology, May 2021. https://cset.georgetown.edu/publication/contending-frames/.; on the history of AI narratives see also generally: Cave, Stephen, Kanta Dihal, and Sarah Dillon, eds. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. New York: Oxford University Press, 2020.

→ Expert views on best practices in AGI lab safety and governance.[238]

Predicting future shifts in perceptions of AI among relevant actors given:

→ The spread of ongoing academic conversations concerned about advanced AI risk;[239]

→ The effects of "warning shots,"[240] or other "risk awareness moments";[241]

→ The effect of motivated misinformation or politicized AI risk skepticism.[242]

3.2. Stakeholder trust in AI developers

→ Public trust in different actors to responsibly develop AI;[243]

→ AI-practitioner trust in different actors to responsibly develop AI[244] and Chinese AI researchers' views on the development of "strong AI."[245]

3.3. Default landscape of regulations applied to AI

This work maps the prevailing (i.e., default, "business-as-usual") landscape of regulations that will be applied to AI in the near term. These matter as they will directly affect the development landscape for advanced AI and indirectly bracket the space for any new (AI-specific) governance proposals.[246] This work includes:

→ Existing industry norms and practices applied to AI in areas such as release practices around generative AI systems;[247]

---

[238] Schuett, Jonas, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. 'Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion'. arXiv, 11 May 2023. https://doi.org/10.48550/arXiv.2305.07153.

[239] Krueger, David Scott. 'An Update on Academia vs. Industry (One Year into My Faculty Job)'. AI Alignment Forum, 3 September 2022. https://www.alignmentforum.org/posts/HXxHcRCxR4oHrAsEr/an-update-on-academia-vs-industry-one-year-into-my-faculty. But see also: Bensinger, Rob. 'The Inordinately Slow Spread of Good AGI Conversations in ML'. LessWrong, 21 June 2022. https://www.lesswrong.com/posts/Rkxj7TFxhbm59AKJh/the-inordinately-slow-spread-of-good-agi-conversations-in-ml.

[240] Gabs, Nick. 'Lessons from Three Mile Island for AI Warning Shots'. EA Forum, 26 September 2022. https://forum.effectivealtruism.org/posts/NyCHoZGGw5YssvDJB/lessons-from-three-mile-island-for-ai-warning-shots.

[241] Guest, Oliver. 'Prospects for AI Safety Agreements between Countries'. Rethink Priorities, 14 April 2022. https://rethinkpriorities.org/publications/prospects-for-ai-safety-agreements-between-countries.; Guest, Oliver. '"Risk Awareness Moments" (Rams): A Concept for Thinking about AI Governance Interventions'. EA Forum, 14 April 2023. https://forum.effectivealtruism.org/posts/EcrNFxGszfgcGevtf/risk-awareness-moments-rams-a-concept-for-thinking-about-ai.

[242] Baum, Seth D. 'Countering Superintelligence Misinformation'. *Information* 9, no. 10 (30 September 2018): 244. https://doi.org/10.3390/info9100244. ; Baum, Seth D. 'Superintelligence Skepticism as a Political Tool'. *Information* 9, no. 9 (22 August 2018): 209. https://doi.org/10.3390/info9090209.

[243] Zhang, Baobao, and Allan Dafoe. 'U.S. Public Opinion on the Governance of Artificial Intelligence'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 187–93. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375827.

[244] Zhang, Baobao, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C. Horowitz, and Allan Dafoe. 'Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers'. *Journal of Artificial Intelligence Research* 71 (2 August 2021): 591-666-591–666. https://doi.org/10.1613/jair.1.12895.

[245] Zeng, Yi, and Kang Sun. 'Whether We Can and Should Develop Strong AI: A Survey in China'. Center for Long-term Artificial Intelligence, 12 March 2023. https://long-term-ai.center/research/f/whether-we-can-and-should-develop-strong-artificial-intelligence.

[246] This overlaps with the evaluation of actors' levers of control as well as pathways of influence.

[247] Solaiman, Irene. 'The Gradient of Generative AI Release: Methods and Considerations'. arXiv, 5 February 2023. https://doi.org/10.48550/arXiv.2302.04844.

→ General existing laws and governance regimes which may be extended to or affect AI development, such as anticompetition law;[248] national and international standards;[249] international law norms, treaties, and regimes;[250] and existing global governance institutions.[251]

→ AI-specific governance regimes currently under development, such as:
  → EU: the EU AI Act [252] and the AI Liability Directive,[253] amongst others;

[248] Hua, Shin-Shin, and Haydn Belfield. 'AI & Antitrust: Reconciling Tensions Between Competition Law and Cooperative AI Development'. *Yale Journal of Law and Technology* 23 (Spring 2021): 127. https://yjolt.org/ai-antitrust-reconciling-tensions-between-competition-law-and-cooperative-ai-development

[249] Cihon, Peter. 'Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development'. Technical Report. Oxford: Center for the Governance of AI, Future of Humanity Institute, University of Oxford, April 2019. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.; O'Keefe, Cullen, Jade Leung, and Markus Anderljung. 'How Technical Safety Standards Could Promote TAI Safety'. Effective Altruism Forum, 8 August 2022. https://forum.effectivealtruism.org/posts/zvbGXCxc5jBowCuNX/how-technical-safety-standards-could-promote-tai-safety .

[250] Kunz, Martina, and Seán Ó hÉigeartaigh. 'Artificial Intelligence and Robotization'. In *Oxford Handbook on the International Law of Global Security*, edited by Robin Geiss and Nils Melzer. Oxford University Press, 2021. https://papers.ssrn.com/abstract=3310421.; Kemp, Luke, and Catherine Rhodes. 'The Cartography of Global Catastrophic Governance'. Global Challenges Foundation, 2020. https://globalchallenges.org/the-cartography-of-global-catastrophic-governance/. (pg. 4-6); Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf (pg. 94-118); Garcia, Eugenio V. 'Multilateralism and Artificial Intelligence: What Role for the United Nations?' In *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, 18. Boca Raton: CRC Press, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3779866. See also the overview at: Kunz, Martina. 'Global AI Governance'. Accessed 26 August 2022. https://globalaigov.org/#. For an older account, see also Castel, J.G., and Mathew E. Castel. 'The Road to Artificial Superintelligence - Has International Law a Role to Play?' *Canadian Journal of Law & Technology* 14 (2016). https://ojs.library.dal.ca/CJLT/article/download/7211/6256.

[251] Schmitt, Lewin. 'Mapping Global AI Governance: A Nascent Regime in a Fragmented Landscape'. *AI and Ethics*, 17 August 2021. https://doi.org/10.1007/s43681-021-00083-y.; see also Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. (pg. 94-121). Roberts, Huw, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. 'Global AI Governance: Barriers and Pathways Forward'. SSRN Scholarly Paper. Rochester, NY, 29 September 2023. https://doi.org/10.2139/ssrn.4588040.

[252] The literature of commentaries is vast, but for a sample see: Veale, Michael, and Frederik Zuiderveen Borgesius. 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach'. *Computer Law Review International* 22, no. 4 (1 August 2021): 97–112. https://doi.org/10.9785/cri-2021-220402.; Schuett, Jonas. 'Risk Management in the Artificial Intelligence Act'. *European Journal of Risk Regulation*, 8 February 2023, 1–19. https://doi.org/10.1017/err.2023.1. Almada, Marco, and Nicolas Petit. 'The EU AI Act: Between Product Safety and Fundamental Rights'. SSRN Scholarly Paper. Rochester, NY, 20 December 2022. https://doi.org/10.2139/ssrn.4308072.

[253] European Commission. 'Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive)'. European Commission, 28 September 2022. https://ec.europa.eu/info/sites/default/files/1_1_197605_prop_dir_ai_en.pdf.; The literature of commentaries is vast, but for a sample see: Ziosi, Marta, Jakob Mökander, Claudio Novelli, Federico Casolari, Mariarosaria Taddeo, and Luciano Floridi. 'The EU AI Liability Directive: Shifting the Burden From Proof to Evidence'. SSRN Scholarly Paper. Rochester, NY, 6 June 2023. https://doi.org/10.2139/ssrn.4470725. Madiega, Tambiama. 'Artificial Intelligence Liability Directive'. European Parliamentary Research Service - Scientific Foresight Unit (STOA), February 2023. See also: Hacker, Philipp. 'The European AI Liability Directives -- Critique of a Half-Hearted Approach and Lessons for the Future'. arXiv, 23 January 2023. https://doi.org/10.48550/arXiv.2211.13960.

→ US: the US AI policy agenda,[254] such as various federal legislative proposals relating to generative AI,[255] or President Biden's executive order,[256] amongst others.;

→ International: such as the 2019 OECD AI Principles (nonbinding);[257] the 2021 UNESCO Recommendation on the Ethics of Artificial Intelligence (nonbinding);[258] the 2023 G7 Hiroshima guidelines (nonbinding);[259] and the Council of Europe's draft (framework) Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law (potentially binding),[260] amongst others.

3.4. Prevailing barriers to effective AI governance

→ Definitional complexities of AI as target for regulation;[261]

→ Potential difficulties around building global consensus given geopolitical stakes and tensions;[262]

[254] Schiff, Daniel S. 'Looking through a Policy Window with Tinted Glasses: Setting the Agenda for U.S. AI Policy'. *Review of Policy Research* n/a, no. n/a. Accessed 1 February 2023. https://doi.org/10.1111/ropr.12535.

[255] Lenhart, Anna. 'Roundup of Federal Legislative Proposals That Pertain to Generative AI'. Tech Policy Press, 21 April 2023. https://techpolicy.press/roundup-of-federal-legislative-proposals-that-pertain-to-generative-ai/.

[256] Biden, Joseph R. 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'. The White House, 30 October 2023. 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/.; and see Office of Management and Budget. 'OMB Releases Implementation Guidance Following President Biden's Executive Order on Artificial Intelligence'. The White House, 1 November 2023. https://www.whitehouse.gov/omb/briefing-room/2023/11/01/omb-releases-implementation-guidance-following-president-bidens-executive-order-on-artificial-intelligence/.

[257] OECD. 'Recommendation of the Council on Artificial Intelligence'. OECD Legal Instruments - OECD/LEGAL/0449, 22 May 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. But for discussion of the limited implementation challenges, see: OECD. 'State of Implementation of the OECD AI Principles: Insights from National AI Policies'. OECD Digital Economy Papers. OECD, 2021. https://www.oecd-ilibrary.org/content/paper/1cd40c44-en.

[258] UNESCO. 'Recommendation on the Ethics of Artificial Intelligence', 23 November 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[259] 'Hiroshima Process International Code of Conduct for Advanced AI Systems', 30 October 2023. https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems.

[260] Committee on Artificial Intelligence (CAI). 'Revised Zero Draft [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law'. Council of Europe, 6 January 2023. https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f.; for commentary, see: Breuer, Marten. 'The Council of Europe as an AI Standard Setter'. *Verfassungsblog* (blog), 4 April 2022. https://verfassungsblog.de/the-council-of-europe-as-an-ai-standard-setter/.

[261] Scherer, Matthew U. 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies'. *Harvard Journal of Law & Technology*, no. 2 (Spring 2016). http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf.; Schuett, Jonas. 'Defining the Scope of AI Regulations'. *Law, Innovation and Technology* 0, no. 0 (3 March 2023): 1–23. https://doi.org/10.1080/17579961.2023.2184135. Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf.

[262] Trager, Robert F. 'The Security Governance Challenge of Emerging Technologies'. *Orbis* 66, no. 4 (2022): 536–50. https://doi.org/10.1016/j.orbis.2022.08.008.

→ Potential difficulty around building civil society consensus given outstanding disagreements and tensions between different expert communities;[263]

→ Potential challenges around cultivating sufficient state capacity to effectively implement and enforce AI legislation.[264]

3.5. Effects of AI systems on tools of governance

Predicting the impact of future technologies on governance and the ways these could shift the possibility frontier of what kind of regimes will be politically viable and enforceable:

→ Effects of AI on general cooperative capabilities;[265]

→ Effects of AI on international law creation and enforcement;[266]

[263] See Park, Peter S., and Max Tegmark. 'Divide-and-Conquer Dynamics in AI-Driven Disempowerment'. arXiv, 9 October 2023. https://doi.org/10.48550/arXiv.2310.06009. But for responses, see also Sætra, Henrik Skaug, and John Danaher. 'Resolving the Battle of Short- vs. Long-Term AI Risks'. *AI and Ethics*, 4 September 2023. https://doi.org/10.1007/s43681-023-00336-y. And Price, Huw, and Matthew Connelly. 'AI Governance Must Deal with Long-Term Risks as Well'. *Nature* 622, no. 7981 (3 October 2023): 31–31. https://doi.org/10.1038/d41586-023-03117-z. Price, Huw, and Matthew Connolly. 'Nature and the Machines'. arXiv, 23 July 2023. https://doi.org/10.48550/arXiv.2308.04440. Brauner, Jan, and Alan Chan. 'AI's Long-Term Risks Shouldn't Distract From Present Risks'. *TIME*, 10 August 2023. https://time.com/6303127/ai-future-danger-present-harms/. And see previous arguments including: Stix, Charlotte, and Matthijs M. Maas. 'Bridging the Gap: The Case for an "Incompletely Theorized Agreement" on AI Policy'. *AI and Ethics* 1, no. 3 (15 January 2021): 261–71. https://doi.org/10.1007/s43681-020-00037-w; Prunkl, Carina, and Jess Whittlestone. 'Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 138–43. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375803.; Cave, Stephen, and Seán S. Ó hÉigeartaigh. 'Bridging Near- and Long-Term Concerns about AI'. *Nature Machine Intelligence* 1, no. 1 (January 2019): 5–6. https://doi.org/10.1038/s42256-018-0003-2.; Baum, Seth D. 'Reconciliation between Factions Focused on Near-Term and Long-Term Artificial Intelligence'. *AI & SOCIETY* 33, no. 4 (2018): 565–72. https://doi.org/10.1007/s00146-017-0734-3.

[264] Lawrence, Christie, Isaac Cui, and Daniel Ho. 'The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies'. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 606–52. AIES '23. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3600211.3604701.

[265] Dafoe, Allan, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 'Cooperative AI: Machines Must Learn to Find Common Ground'. *Nature* 593, no. 7857 (May 2021): 33–36. https://doi.org/10.1038/d41586-021-01170-0.

[266] Deeks, Ashley. 'High-Tech International Law'. *George Washington Law Review* 88 (2020): 575–653. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531976 ; Maas, Matthijs M. 'International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order'. *Melbourne Journal of International Law* 20, no. 1 (2019): 29–56. https://law.unimelb.edu.au/__data/assets/pdf_file/0005/3144308/Maas.pdf ; Maas, Matthijs M. 'AI, Governance Displacement, and the (De)Fragmentation of International Law'. In *ISA Annual Convention*, 2021. https://www.cser.ac.uk/resources/ai-governance-displacement-and-defragmentation-international-law/.

→ Effects of AI on arms control monitoring.[267]

4. Other lenses on the advanced AI governance problem

Other work aims to derive key strategic lessons for advanced AI governance, not by aiming to empirically map or estimate first-order facts about the key (technical, deployment, or governance) strategic parameters, but rather by drawing indirect (empirical, strategic, and/or normative) lessons from abstract models, historical cases, and/or political theory.

4.1. Lessons derived from theory

Work characterizing the features of advanced AI technology and of its governance challenge, drawing on existing literatures or bodies of theory:

Mapping clusters and taxonomies of AI's governance problems**:**

→ AI creating distinct types of risk deriving from (1) accidents, (2) misuse, and (3) structure;[268]

→ AI creating distinct problem logics across domains: (1) ethical challenges, (2) safety risks, (3) security threats, (4) structural shifts, (5) common goods, and (6) governance disruption;[269]

→ AI driving four risk clusters: (1) inequality, turbulence, and authoritarianism; (2) great-power war; (3) the problems of control, alignment, and political order; and (4) value erosion from competition.[270]

Mapping the political features of advanced AI technology:

---

[267] Mittelsteadt, Matthew. 'AI Verification: Mechanisms to Ensure AI Arms Control Compliance'. Center for Security and Emerging Technology, February 2021. https://live-cset-georgetown.pantheonsite.io/research/ai-verification/.; see more generally work beyond the community, such as: Vaynman, Jane. 'Better Monitoring and Better Spying: The Implications of Emerging Technology for Arms Control'. *Texas National Security Review* 4, no. 4 (23 September 2021). https://tnsr.org/2021/09/better-monitoring-and-better-spying-the-implications-of-emerging-technology-for-arms-control/.; Reinhold, Thomas, and Niklas Schörnig. *Armament, Arms Control and Artificial Intelligence: The Janus-Faced Nature of Machine Learning in the Military Realm*. Springer Nature, 2022. https://link.springer.com/book/10.1007/978-3-031-11043-6; Lück, Nico. *Machine Learning-Powered Artificial Intelligence in Arms Control*. PRIF Report 2019, 8. Frankfurt am Main: Peace Research Institute Frankfurt, 2019. https://www.hsfk.de/publikationen/publikationssuche/publikation/machine-learning-powered-artificial-intelligence-in-arms-control.; Schörnig, Niklas. 'AI for Arms Control : How Artificial Intelligence Can Foster Verification and Support Arms Control'. Peace Research Institute Frankfurt, 2022. https://doi.org/10.48809/PRIFSPOT2201. See also Cox, Jessica, and Heather Williams. 'The Unavoidable Technology: How Artificial Intelligence Can Strengthen Nuclear Stability'. *The Washington Quarterly* 44, no. 1 (2 January 2021): 69–85. https://doi.org/10.1080/0163660X.2021.1893019. Pg. 77-79 (on AI applications in arms control).

[268] Zwetsloot, Remco, and Allan Dafoe. 'Thinking About Risks From AI: Accidents, Misuse and Structure'. Lawfare, 11 February 2019. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.

[269] Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. (Chapter 4).; Maas, Matthijs M. 'Aligning AI Regulation to Sociotechnical Change'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang. Oxford University Press, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.22.

[270] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press, 2022. https://docs.google.com/document/d/e/2PACX-1vQOQ0EBIaEu_LaJqWvdPKu8xlmrOCM6h6gq7eFHnN0Y2GPYoodQjLeilxQ8SUwnbVThXc0k_jCIsCX1/pub

→ AI as general-purpose technology, highlighting radical impacts on economic growth, disruption to existing socio-political relations, and potential for backlash and social conflict;[271]

→ AI as industry-configured general-purpose tech (low fixed costs and private sector dominance), highlighting challenges of rapid proliferation (compared to "prestige," "public," or "strategic" technologies);[272]

→ AI as information technology, highlighting challenges of increasing returns to scale driving greater income inequality, impacts on broad collective identities as well as community fragmentation, and increased centralization of (cybernetic) control;[273]

→ AI as intelligence technology, highlighting challenges of bias, alignment, and control of the principal over the agent;[274]

→ AI as regulation-resistant technology, rendering coordinated global regulation difficult.[275]

Mapping the structural features of the advanced AI governance challenge:

[271] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press, 2022. https://docs.google.com/document/d/e/2PACX-1vQOQ0EBIaEu_LaJqWvdPKu8xlmrOCM6h6gq7eFHnN0Y2GPYoodQjLeilxQ8SUwnbVThXc0k_jCIsCX1/pub. For an argument that current large language models may have reached the level of performance to be GPTs, see: Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 'GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models'. arXiv, 19 March 2023. https://doi.org/10.48550/arXiv.2303.10130. For a critical counter-argument, claiming that AI is better understood not as GPT, but through the "Large Technical Systems (LTS)" lens, see Vannuccini, Simone, and Ekaterina Prytkova. 'Artificial Intelligence's New Clothes? From General Purpose Technology to Large Technical System'. 7 April 2021. https://doi.org/10.2139/ssrn.3860041.

[272] Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. pg 77-78; drawing on: Drezner, Daniel W. 'Technological Change and International Relations'. *International Relations* 33, no. 2 (1 June 2019): 286–303. https://doi.org/10.1177/0047117819834629.

[273] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press, 2022. https://docs.google.com/document/d/e/2PACX-1vQOQ0EBIaEu_LaJqWvdPKu8xlmrOCM6h6gq7eFHnN0Y2GPYoodQjLeilxQ8SUwnbVThXc0k_jCIsCX1/pub.

[274] Ibid.

[275] Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. Pg. 263-264; drawing on frameworks presented in: Crootof, Rebecca. 'Jurisprudential Space Junk: Treaties and New Technologies'. In *Resolving Conflicts in the Law*, edited by Chiara Giorgetti and Natalie Klein, 106–29, 2019. https://brill.com/view/book/edcoll/9789004316539/BP000015.xml.; Watts, Sean. 'Regulation-Tolerant Weapons, Regulation-Resistant Weapons and the Law of War'. *International Law Studies* 91 (2015): 83. https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1411&context=ils ; Watts, Sean. 'Autonomous Weapons: Regulation Tolerant or Regulation Resistant?' SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 9 October 2015. https://doi.org/10.2139/ssrn.2681283.

→ In terms of its intrinsic coordination challenges: as a global public good,[276] as a collective action problem,[277] and as a matter of "existential security";[278]

→ In terms of its difficulty of successful resolution: as a wicked problem[279] and as a challenge akin to "racing through a minefield";[280]

[276] AI Impacts. 'Friendly AI as a Global Public Good'. AI Impacts, 8 August 2016. https://aiimpacts.org/friendly-ai-as-a-global-public-good/. See also Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. (Chapter 2.2).

[277] Neufville, Robert de, and Seth D. Baum. 'Collective Action on Artificial Intelligence: A Primer and Review'. *Technology in Society* 66 (1 August 2021): 101649. https://doi.org/10.1016/j.techsoc.2021.101649; Askell, Amanda, Miles Brundage, and Gillian Hadfield. 'The Role of Cooperation in Responsible AI Development', 10 July 2019, 23. https://arxiv.org/abs/1907.04534

[278] See: Sears, Nathan Alexander. 'Existential Security: Towards a Security Framework for the Survival of Humanity'. *Global Policy* 11, no. 2 (2020): 255–66. https://doi.org/10.1111/1758-5899.12800.; Sears, Nathan Alexander. 'International Politics in the Age of Existential Threats'. *Journal of Global Security Studies*, 18 June 2020, 1–23. https://doi.org/10.1093/jogss/ogaa027.

[279] Gruetzemacher, Ross. 'Rethinking AI Strategy and Policy as Entangled Super Wicked Problems', 6. AIES 2018; New Orleans, 2018. http://www.rossgritz.com/wp-content/uploads/2018/11/aies_gruetzemacher_revisions.pdf ; Liu, Hin-Yan, and Matthijs M. Maas. '"Solving for X?" Towards a Problem-Finding Framework to Ground Long-Term Governance Strategies for Artificial Intelligence'. *Futures* 126 (1 February 2021): 22. https://doi.org/10.1016/j.futures.2020.102672.

[280] Karnofsky, Holden. 'Racing through a Minefield: The AI Deployment Problem'. Cold Takes, 22 December 2022. https://www.cold-takes.com/racing-through-a-minefield-the-ai-deployment-problem/.

→ In terms of its strategic dynamics: as a *technology race*,[281] whether motivated by security concerns or by prestige motivations,[282] or as an arms race[283] (but see also critiques of the arms race framing on definitional grounds,[284] on empirical grounds,[285] and on grounds of rhetorical or framing risks[286]);

→ In terms of its politics and power dynamics: as a political economy problem.[287]

Identifying design considerations for international institutions and regimes, from:

→ General theory on the rational design of international institutions;[288]

[281] Han, The Anh, Luis Moniz Pereira, and Tom Lenaerts. 'Modelling and Influencing the AI Bidding War: A Research Agenda', 2019. http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_28.pdf.; Stafford, Eoghan, Robert Trager, and Allan Dafoe. 'International Strategic Dynamics of Risky Technology Races', June 2021. https://www.academia.edu/49586612/International_Strategic_Dynamics_of_Risky_Technology_Races

[282] Barnhart, Joslyn. 'Emerging Technologies, Prestige Motivations and the Dynamics of International Competition', January 2022, 56. https://www.governance.ai/research-paper/emerging-technologies-prestige-motivations-and-the-dynamics-of-international-competition

[283] Shulman, Carl. 'Arms Control and Intelligence Explosions', 6. Bellaterra, Spain, 2009. https://intelligence.org/files/ArmsControl.pdf.; Armstrong, Stuart, Nick Bostrom, and Carl Shulman. 'Racing to the Precipice: A Model of Artificial Intelligence Development'. *AI & Society* 31, no. 2 (2016): 201–6. https://link.springer.com/article/10.1007/s00146-015-0590-y

[284] Scharre, Paul. 'Debunking the AI Arms Race Theory'. *Texas National Security Review* 4, no. 3 (28 June 2021). https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/.; Roff, Heather M. 'The Frame Problem: The AI "Arms Race" Isn't One'. *Bulletin of the Atomic Scientists* 0, no. 0 (26 April 2019): 1–4. https://doi.org/10.1080/00963402.2019.1604836.

[285] Kania, Elsa B. 'Technological Entanglement: Cooperation, Competition and the Dual-Use Dilemma in Artificial Intelligence'. Policy Brief. Australian Strategic Policy Institute, 2018. https://s3-ap-southeast-2.amazonaws.com/ad-aspi/2018-07/Tech-Entanglemen_PolicyBrief_20180702-v2.pdf?7BahbUgNHCY1umz4PCrLOEdBJUrjULCg. ; Bryson, Joanna J., and Helena Malikova. 'Is There an AI Cold War?' *Global Perspectives* 2, no. 1 (28 June 2021): 24803. https://doi.org/10.1525/gp.2021.24803.; see also Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press, 2022. https://docs.google.com/document/d/e/2PACX-1vQOQ0EBIaEu_LaJqWvdPKu8xlmrOCM6h6gq7eFHnN0Y2GPYoodQjLeilxQ8SUwnbVThXc0k_jCIsCX1/pub. ("At present the 'arms' modifier is largely literally off-point, since most of the geopolitical activity in AI is not about weapons per se, but is instead about supply chains, infrastructure, industrial base, strategic industries, scientific capability, and prestige achievements").

[286] Cave, Stephen, and Seán S. ÓhÉigeartaigh. 'An AI Race for Strategic Advantage: Rhetoric and Risks'. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 36–40. New Orleans LA USA: ACM, 2018. https://doi.org/10.1145/3278721.3278780.; Belfield, Haydn. 'Are You Really in a Race? The Cautionary Tales of Szilárd and Ellsberg'. EA Forum, 2022. https://forum.effectivealtruism.org/posts/cXBznkfoPJAjacFoT/are-you-really-in-a-race-the-cautionary-tales-of-szilard-and.

[287] Kemp, Luke. 'Agents of Doom: Who Is Creating the Apocalypse and Why'. *BBC Future*, 26 October 2021. https://www.bbc.com/future/article/20211014-agents-of-doom-who-is-hastening-the-apocalypse-and-why.

[288] See generally: Koremenos, Barbara, Charles Lipson, and Duncan Snidal. 'The Rational Design of International Institutions'. *International Organization* 55, no. 4 (ed 2001): 761–99. https://doi.org/10.1162/002081801317193592.

→ Theoretical work on the orchestration and organization of regime complexes of many institutions, norms, conventions, etc.[289]

4.2. Lessons derived from models and wargames

Work to derive or construct abstract models for AI governance in order to gather lessons from these for understanding AI systems' proliferation and societal impacts. This includes models of:

→ International strategic dynamics in risky technology races,[290] and theoretical models of the role of information sharing,[291] agreement, or incentive modeling;[292]

→ AI competition and whether and how AI safety insights will be applied under different AI safety-performance tradeoffs,[293] including collaboration on safety as a social dilemma[294] and models of how compute pricing factors affect agents' spending on safety ("safety tax") meant to reduce the danger from the new technology;[295]

→ The offense-defense balance of increasing investments in technologies;[296]

→ The offense-defense balance of scientific knowledge in AI with potential for misuse;[297]

---

[289] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890.; Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Should Artificial Intelligence Governance Be Centralised?: Design Lessons from History'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–34. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375857. See also generally: Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. (Chapter 7). Tallberg, Jonas, Eva Erman, Markus Furendal, Johannes Geith, Mark Klamberg, and Magnus Lundgren. 'The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research'. *International Studies Review* 25, no. 3 (1 September 2023): viad040. https://doi.org/10.1093/isr/viad040.; Roberts, Huw, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. 'Global AI Governance: Barriers and Pathways Forward'. SSRN Scholarly Paper. Rochester, NY, 29 September 2023. https://doi.org/10.2139/ssrn.4588040.

[290] Stafford, Eoghan, Robert F Trager, and Allan Dafoe. 'Safety Not Guaranteed: International Strategic Dynamics of Risky Technology Races', June 2022, 31.

[291] Emery-Xu, Nicholas, Andrew Park, and Robert Trager. 'Uncertainty, Information, and Risk in International Technology Races', June 2022. https://drive.google.com/file/d/18j_wnA4HDMA3ofclLcfpgyV-0INMn1ZW/view?usp=sharing&.

[292] Han, The Anh, Luis Moniz Pereira, and Tom Lenaerts. 'Modelling and Influencing the AI Bidding War: A Research Agenda'. In *AAAI/ACM Conference on AI, Ethics, and Society*, 5–11, 27 January 2019. https://doi.org/10.1145/3306618.3314265, also available at http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_28.pdf. Han, The Anh, Luis Moniz Pereira, Tom Lenaerts, and Francisco C. Santos. 'Mediating Artificial Intelligence Developments through Negative and Positive Incentives'. *ArXiv:2010.00403 [Nlin, q-Bio]*, 1 October 2020. http://arxiv.org/abs/2010.00403.

[293] Bova, Paolo, Jonas Emanuel Müller, Tanja Rüegg, and Robert Trager. 'Announcing the SPT Model Web App for AI Governance'. EA Forum, 4 August 2022. https://forum.effectivealtruism.org/posts/c73nsggC2GQE5wBjq/announcing-the-spt-model-web-app-for-ai-governance.; See model at https://spt.modelingcooperation.com/; See also Robert Trager, Paolo Bova, Nicholas Emery-Xu, Eoghan Stafford, and Allan Dafoe, "Welfare Implications of Safety-Performance Tradeoffs in AI Safety Research", Working paper, August 2022.

[294] Han, The Anh, Francisco C. Santos, Luís Moniz Pereira, and Tom Lenaerts. 'A Regulation Dilemma in Artificial Intelligence Development'. MIT Press, 2021. https://doi.org/10.1162/isal_a_00385.

[295] Jensen, Mckay, Nicholas Emery-Xu, and Robert Trager. 'Industrial Policy for Advanced AI: Compute Pricing and the Safety Tax'. arXiv, 22 February 2023. https://doi.org/10.48550/arXiv.2302.11436.

[296] Garfinkel, Ben, and Allan Dafoe. 'How Does the Offense-Defense Balance Scale?' *Journal of Strategic Studies* 42, no. 6 (19 September 2019): 736–63. https://doi.org/10.1080/01402390.2019.1631810.

[297] Shevlane, Toby, and Allan Dafoe. 'The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?' In *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020. http://arxiv.org/abs/2001.00463.

→ Lessons from the "epistemic communities" lens, on how coordinated expert networks can shape policy;[298]

→ Lessons from wargames and role-playing exercises.[299]

4.3. Lessons derived from history

Work to identify and study relevant historical precedents, analogies, or cases and to derive lessons for (AI) governance.[300] This includes studies where historical cases have been directly applied to advanced AI governance as well as studies where the link has not been drawn but which might nevertheless offer productive insights for the governance of advanced AI.

Lessons from the history of technology development and spread

Historical cases that (potentially) provide insights into when, why, and how new technologies are pursued and developed—and how they subsequently (fail to) spread.

Historical rationales for technology pursuit and development

Historical rationales for actors pursuing large-scale scientific or technology development programs:

→ Development of major transformative technologies during wartime: US development of the atom bomb;[301]

→ Pursuit of strategically valuable megaprojects: the Apollo Program and the Manhattan Project;[302]

[298] Pulver, Tobias. 'Shaping Policy as Experts: An Epistemic Community for (Transformative) AI Governance?' 2019. https://docs.google.com/document/d/1h7YHlp44kANhXPo8zJdr5ea1mttF8E5l-pJ_w8v_quE/edit?usp=drive_web&ouid=107201564093427841585&

[299] Avin, Shahar, Ross Gruetzemacher, and James Fox. 'Exploring AI Futures Through Role Play'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8–14. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375817.; see more broadly the review of methodologies in Avin, Shahar. 'Exploring Artificial Intelligence Futures'. *AIHumanities*, 17 January 2019. http://aihumanities.org/en/journal/past-issues/?board_name=Enjournal&search_field=fn_title&search_text=Exploring%20&vid=15.

[300] This list is obviously not exhaustive. It lists cases that have been identified, flagged, or studied by researchers in the field; however, there are many additional possible cases. In the discussion of "historical analogies" of each perspective, I will suggest a number of additional plausible historical cases that could yield valuable lessons, insights, or support to a given perspective as well as counterexamples that highlight potential failure modes or barriers to be overcome.

[301] Ord, Toby. 'Lessons from the Development of the Atomic Bomb'. Center for the Governance of AI, November 2022. https://www.governance.ai/research-paper/lessons-atomic-bomb-ord. (exploring insights in terms of the scientific, engineering and political prerequisites for making the atomic bomb; the labor and money invested; the role and efficacy of secrecy in various national programmes; the role of spying; the ability of scientists to provide decision makers with useful estimates of the cost and effects of an atomic bomb; US decision-making about whether and how to use the bomb; the effects of the atomic bombings of Japan; the subsequent efforts of atomic scientists to control the development and use of the technology; the impact of individual actors on the development of atomic weapons; and how scientists managed the potential existential risk of nuclear weapons igniting the atmosphere).

[302] Levin, John-Clark, and Matthijs M. Maas. 'Roadmap to a Roadmap: How Could We Tell When AGI Is a "Manhattan Project" Away?', 7. Santiago de Compostela, Spain, 2020. http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_11.pdf.

→ Technologies pursued for prestige reasons: Ming Dynasty treasure fleets,[303] the US/USSR space race,[304] and the French nuclear weapons program;[305]

→ Risk of races being started by possibly incorrect perceptions that a rival is actively pursuing a technology: the Manhattan Project (1939–1945), spurred by the Einstein Letter; the "missile gap" project to build up a US ICBM capability (1957–1962).[306]

Historical strategies of deliberate large-scale technology development projects

Historical strategies for unilateral large-scale technology project development:

→ Crash recruitment and resource allocation for a large strategic program: "Operation Paperclip," the post-WWII effort to recruit 1,600 German scientists and engineers, fast-tracking the US space program as well as several programs aimed at other Cold War weapons of mass destruction;[307]

→ Different potential strategies for pursuing advanced strategic technologies: the distinct nuclear proliferation strategies ("hedging, sprinting, sheltered pursuit, hiding") taken by different countries in pursuing nuclear weapons;[308]

→ Government-industry collaborations to boost development of strategic technologies: the 1980's SEMATECH collaborative research consortium to boost the US semiconductor industry;[309]

→ Nations achieving early and sustained unilateral leads in developing key strategic technologies: the US program to develop stealth aircraft;[310]

→ Surprisingly rapid leaps from the political decision to run a big technology program to the achievement: Apollo 8 (134 days between NASA decision to go to the moon and launch),[311] UAE's "Hope" Mars mission (set up its space agency UAESA in 2014, was only able to design its own satellite (KhalifaSat) in 2018, and launched its "Hope" Mars Mission in July 2020, less than six years

---

[303] Musgrave, Paul, and Daniel H. Nexon. 'Defending Hierarchy from the Moon to the Indian Ocean: Symbolic Capital and Political Dominance in Early Modern China and the Cold War'. *International Organization* 72, no. 3 (ed 2018): 591–626. https://doi.org/10.1017/S0020818318000139.

[304] Ibid. and Barnhart, Joslyn. 'Emerging Technologies, Prestige Motivations and the Dynamics of International Competition', January 2022, 56. https://www.governance.ai/research-paper/emerging-technologies-prestige-motivations-and-the-dynamics-of-international-competition

[305] Sagan, Scott D. 'Why Do States Build Nuclear Weapons?: Three Models in Search of a Bomb'. *International Security* 21, no. 3 (1996): 54–86. http://www.jstor.org/stable/2539273

[306] Belfield, Haydn. 'Are You Really in a Race? The Cautionary Tales of Szilárd and Ellsberg'. EA Forum, 2022. https://forum.effectivealtruism.org/posts/cXBznkfoPJAjacFoT/are-you-really-in-a-race-the-cautionary-tales-of-szilard-and .

[307] Crim, Brian E. Our Germans: Project Paperclip and the National Security State. Illustrated edition. Baltimore: Johns Hopkins University Press, 2018.; Haleas, Diane, and Matthew Miller. 'Session B-3: Operation Paperclip and the Rise of Weapons of Mass Destruction'. Professional Learning Day, 4 March 2016. https://digitalcommons.imsa.edu/proflearningday/2016/history/6.

[308] Narang, Vipin. *Seeking the Bomb: Strategies of Nuclear Proliferation*. Princeton Studies in International History and Politics. Princeton University Press, 2022. https://press.princeton.edu/books/paperback/9780691172620/seeking-the-bomb.

[309] Forero, Felipe Calero, and Robert Trager. 'The History of Sematech and Lessons for State-Sponsored Industry Cooperation in AI', 2023.

[310] Westwick, Peter. 'Lessons from Stealth for Emerging Technologies'. Center for Security and Emerging Technology (blog), March 2021. https://cset.georgetown.edu/publication/lessons-from-stealth-for-emerging-technologies/.

[311] NASA. 'The Apollo Spacecraft - A Chronology. Vol. IV. Part 2 (1968 Aug/Sep)'. NASA Special Publication, 1969. https://www.hq.nasa.gov/office/pao/History/SP-4009/v4p2n.htm. ; as mentioned in: Collison, Patrick. 'Fast'. Accessed 1 August 2022. https://patrickcollison.com/fast.

after establishment),[312] and various other examples including BankAmericard (90 days), P-80 Shooting Star (first USAF jet fighter) (143 days), Marinship (197 days), The Spirit of St. Louis (60 days), the Eiffel Tower (2 years and 2 months), Treasure Island, San Francisco (~2 years), the Alaska Highway (234 days), Disneyland (366 days), the Empire State Building (410 days), Tegel Airport and the Berlin Airlift (92 days),[313] the Pentagon (491 days), Boeing 747 (930 days), the New York Subway (4.7 years), TGV (1,975 days), USS Nautilus (first nuclear submarine) (1,173 days), JavaScript (10 days), Unix (21 days), Xerox Alto (first GUI-oriented computer) (4 months), iPod (290 days), Amazon Prime (6 weeks), Git (17 days), and COVID-19 vaccines (3-45 days).[314]

Historical strategies for joint or collaborative large-scale technology development:

- → International "big science" collaborations: CERN, ITER, International Space Station, Human Genome Project,[315] and attempted collaborations on Apollo-Soyuz between the US and Soviet space programs.[316]

Historical instances of sudden, unexpected technological breakthroughs

Historical cases of rapid, historically discontinuous breakthroughs in technological performance on key metrics:

- → "Large robust discontinuities" in historical technology performance trends:[317]
  - → the Pyramid of Djoser (2650 BC—structure height trends);
  - → the SS Great Eastern (1858—ship size trends);
  - → the first and second telegraphs (1858, 1866—speed of sending a message across the Atlantic Ocean);
  - → the first nonstop transatlantic flight (1919—speed of passenger or military payload travel);
  - → first nuclear weapons (1945—relative effectiveness of explosives);
  - → first ICBM (1958—average speed of military payload);

[312] Dowling, Stephen. 'How the UAE Got a Spacecraft to Mars – on the First Try'. BBC Future, 19 December 2022. https://www.bbc.com/future/article/20221206-how-the-uae-got-a-spacecraft-to-mars-on-the-first-try.
[313] Collison, Patrick. 'Fast'. Accessed 1 August 2022. https://patrickcollison.com/fast.
[314] Ibid. Note that the precise timeline on which different COVID-19 vaccines were developed varied: Moderna took 65 days from receiving the genetic sequence of the coronavirus to designing the vaccine, demonstrating its efficacy in vitro and in animals, and starting the first human trial. However, it took 270 additional days for the vaccine to be approved by the FDA under emergency use authorization. Więcek, Witold. 'From Warp Speed to 100 Days'. *Asterisk*, 2023. https://asteriskmag.com/issues/04/from-warp-speed-to-100-days.
[315] Robinson, Mark. 'Big Science Collaborations; Lessons for Global Governance and Leadership'. *Global Policy* n/a, no. n/a (2020). https://doi.org/10.1111/1758-5899.12861. Robinson, Mark. 'The CERN Community; A Mechanism for Effective Global Collaboration?' *Global Policy*, 18 November 2018. https://doi.org/10.1111/1758-5899.12608. (discussing CERN, ITER, and ISS); see also: Kerry, Cameron F, Joshua P Meltzer, and Andrea Renda. 'AI Cooperation on the Ground: AI Research and Development on a Global Scale'. Brookings Institute & Forum for Cooperation on Artificial Intelligence (FCAI), October 2022. https://www.brookings.edu/wp-content/uploads/2022/11/FCAI-October-2022.pdf. Appendix. (discussing the governance and finance arrangements of the HGP, ISS, and CERN, to derive lessons for AI).
[316] Krige, John, Angelina Long Callahan, and Ashok Maharaj. 'Sustaining Soviet-American Collaboration, 1957–1989'. In *NASA in the World: Fifty Years of International Collaboration in Space*, edited by John Krige, Angelina Long Callahan, and Ashok Maharaj, 127–51. Palgrave Studies in the History of Science and Technology. New York: Palgrave Macmillan US, 2013. https://doi.org/10.1057/9781137340931_7. I thank Christian Ruhl for this suggestion.
[317] Grace, Katja. 'Discontinuous Progress in History: An Update'. AI Impacts, 13 April 2020. https://aiimpacts.org/discontinuous-progress-in-history-an-update/. (defining such "large robust discontinuities" as events which "abruptly and clearly contributed more to progress on some technological metric than another century would have seen on the previous trend").

→ the discovery of YBa2Cu3O7 as a superconductor (1987—warmest temperature of superconduction).[318]

→ "Bolt-from-the-blue" technology breakthroughs that were held to be unlikely or impossible even shortly before they happened: Invention of flight;[319] of penicillin, nuclear fission, nuclear bombs, or space flight;[320] of internet hyperlinks and effective internet search.[321]

Historical patterns in technological proliferation and take-up

Historical cases of technological proliferation and take-up:[322]

→ Patterns in the development, dissemination and impacts of major technological advancements: flight, the telegraph, nuclear weapons, the laser, penicillin, the transistor, and others;[323]

→ Proliferation and penetration rates of other technologies in terms of time between invention and widespread use: steam engine (80 years), electricity (40 years), IT (20 years),[324] and mobile phones;

→ Role of state "diffusion capacity" in supporting the diffusion or wide adoption of new innovations: the US in the Second Industrial Revolution and the Soviet Union in the early postwar period;[325]

---

[318] Ibid. (In addition, they also identify five "moderate robust discontinuities" (events that suddenly contribute around 10–100 years of progress of previous trends).

[319] Yudkowsky, Eliezer. 'There's No Fire Alarm for Artificial General Intelligence'. *Machine Intelligence Research Institute* (blog), 14 October 2017. https://intelligence.org/2017/10/13/fire-alarm/. See also Schwartz, Baron. 'Heavier-Than-Air Flight Is Impossible', 4 June 2017. https://www.xaprb.com/blog/flight-is-impossible/.

[320] See examples discussed in: Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. (pg. 62-63, citing sources).

[321] Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. Pg. 62, ftn 169 ("in the first edition of his 2001 book *On the Internet*, Hubert Dreyfus built on his previous critique of AI to argue against the very possibility of searching the internet, claiming that without embodied knowledge, online search would hit an intractable wall. [...] These sections were quietly dropped from the book's Second Edition, published after Google's 2004 IPO."). Drawing on: Dreyfus, Hubert. *On the Internet*. 1st ed. Routledge, 2001. https://www.abebooks.com/first-edition/Internet-Thinking-Action-HUBERT-DREYFUS-Routledge/30612233740/bd.

[322] For work on the broader relevance of technological diffusion patterns, rather than just innovation capabilities, in determining national competitiveness, see: Ding, Jeffrey. 'The Diffusion Deficit in Scientific and Technological Power: Re-Assessing China's Rise'. *Review of International Political Economy* 0, no. 0 (13 March 2023): 1–26. https://doi.org/10.1080/09692290.2023.2173633.

[323] Korzekwa, Rick. 'Observed Patterns around Major Technological Advancements'. AI Impacts, 2 February 2022. https://aiimpacts.org/observed-patterns-around-major-technological-advancements/. For an overview of the underlying 53 case studies, see: AI Impacts. 'Discontinuous Progress Investigation'. *AI Impacts* (blog), 2 February 2015. https://aiimpacts.org/discontinuous-progress-investigation/.

[324] Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. Pg. 79. Drawing on: Gill, Indermit. 'Whoever Leads in Artificial Intelligence in 2030 Will Rule the World until 2100'. *Brookings* (blog), 17 January 2020. https://www.brookings.edu/blog/future-development/2020/01/17/whoever-leads-in-artificial-intelligence-in-2030-will-rule-the-world-until-2100/.; Comin, Diego, and Martí Mestieri. 'If Technology Has Arrived Everywhere, Why Has Income Diverged?' *American Economic Journal: Macroeconomics* 10, no. 3 (July 2018): 137–78. https://doi.org/10.1257/mac.20150175; Comin, Diego, and Martı Mestieri. 'Technology Adoption and Growth Dynamics', 2014, 38. https://pdfs.semanticscholar.org/1f76/728473ee4fb154fe1655a4645c4b43b29358.pdf.; See more generally the discussion of distinct types of technologies and their proliferation profiles, in Drezner, Daniel W. 'Technological Change and International Relations'. *International Relations* 33, no. 2 (1 June 2019): 286–303. https://doi.org/10.1177/0047117819834629.

[325] Ding, Jeffrey. 'The Diffusion Deficit in Scientific and Technological Power: Re-Assessing China's Rise'. *Review of International Political Economy* 0, no. 0 (13 March 2023): 1–26. https://doi.org/10.1080/09692290.2023.2173633.

→ Role of espionage in facilitating critical technology diffusion: early nuclear proliferation[326] and numerous information leaks in modern IT systems;[327]

→ Constrained proliferation of technological insights (even under compromised information security conditions): surprisingly limited track record of bioweapon proliferation: the American, Soviet, Iraqi, South African, and Aum Shinrikyo bioweapon programs ran into a range of problems which resulted in programs that failed if not totally then at least to make effective steps towards weaponization. This suggests that tacit knowledge and organizational conditions can be severely limiting and prevent proliferation even when some techniques are available in the public scientific literature.[328] The (1991–2018) limited success of China in re-engineering US fifth-generation stealth fighters in spite of extensive espionage that included access to blueprints, recruitment of former engineers, and even access to the wreck of a F-117 aircraft that had crashed in Serbia;[329]

→ Various factors contributing to technological delay or restraint with many examples of technologies being slowed or abandoned or having their uptake inhibited, including weapon systems, nuclear power, geoengineering, and genetically modified (GM) crops, as a result of (indirect) regulations, public opposition, and historical contingency;[330]

→ Supply chain evolution of previous general-purpose technologies: studies of railroads, electricity, and cloud computing industries, where supply chains were initially vertically integrated but then evolved into a fully disintegrated natural monopoly structure with a handful of primary "upstream" firms selling services to many "downstream" application sectors.[331]

Lessons from the historical societal impacts of new technologies

Historical cases that (potentially) provide insights into when, why, and how new technologies can have (unusually) significant societal impacts or pose acute risks.

---

[326] Ord, Toby. 'Lessons from the Development of the Atomic Bomb'. Center for the Governance of AI, November 2022. https://www.governance.ai/research-paper/lessons-atomic-bomb-ord.; see also: GAA. 'Nuclear Espionage and AI Governance'. EA Forum, 2021. https://forum.effectivealtruism.org/posts/CKfHDw5Lmoo6jahZD/nuclear-espionage-and-ai-governance-1.

[327] Muelhauser, Luke. 'Example High-Stakes Information Security Breaches [Public]', June 2020. https://docs.google.com/document/d/1_smEDPWDVIaLuZ14Cm7KLHcWx4LkJ0DCTk8bcHjYy_Y/edit#heading=h.hqf76e8phc7g.

[328] Ben Ouagrham-Gormley, Sonia. Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development, 2014. https://doi.org/10.7591/cornell/9780801452888.001.0001. ("The specific organizational, managerial, social, political, and economic conditions necessary for success are difficult to achieve, particularly in covert programs where the need to prevent detection imposes managerial and organizational conditions that conflict with knowledge production."). See also review by Georgia Ray: Ray, Georgia. 'Book Review: Barriers to Bioweapons'. Eukaryote Writes Blog (blog), 30 June 2017. https://eukaryotewritesblog.com/2017/06/30/book-review-barriers/.

[329] Gilli, Andrea, and Mauro Gilli. 'Why China Has Not Caught Up Yet: Military-Technological Superiority and the Limits of Imitation, Reverse Engineering, and Cyber Espionage'. International Security 43, no. 3 (1 February 2019): 141–89. https://doi.org/10.1162/isec_a_00337.

[330] See also: Maas, Matthijs. 'Paths Untaken: The History, Epistemology and Strategy of Technological Restraint, and Lessons for AI'. *Verfassungsblog* (blog), 9 August 2022. https://verfassungsblog.de/paths-untaken/. See also AI Impacts. 'Resisted Technological Temptations Project'. [AI Impacts Wiki], 2023. https://wiki.aiimpacts.org/responses_to_ai/technological_inevitability/incentivized_technologies_not_pursued/resisted_technological_temptations_project.; AI Impacts. 'Incentivized Technologies Not Pursued'. [AI Impacts Wiki], 2023. https://wiki.aiimpacts.org/responses_to_ai/technological_inevitability/incentivized_technologies_not_pursued/start. Heninger, Jeffrey. 'Muddling Along Is More Likely Than Dystopia'. AI Impacts, 20 October 2023. https://blog.aiimpacts.org/p/muddling-along-is-more-likely-than?utm_medium=android.; Heninger, Jeffrey. 'Why Has Geoengineering Been Rejected?' AI Impacts, 2023. http://aiimpacts.org/wp-content/uploads/2023/04/Why-Has-Geoengineering-Been-Rejected.pdf.

[331] Salisbury, Adam. 'How Will the AI Supply Chain Evolve?', 2022. https://docs.google.com/document/d/1s3QGFJ8Ochosksl4JgQCWekJrsY3YFAfGgEiEt6zFpA/edit?usp=sharing&. (pg. 21-25).

Historical cases of large-scale societal impacts from new technologies

Historical cases of large-scale societal impacts from new technologies:[332]

- → Impacts of previous narrowly transformative technologies: impact of nuclear weapons on warfare, and electrification of militaries as driver of "general-purpose military transformation";[333]
- → Impacts of previous general-purpose technologies: general electrification,[334] printing, steam engines, rail transport, motor vehicles, aviation, and computing;[335]
- → Impacts of previous *"revolutionary"* or "*radically transformative"*[336] technologies: domesticated crops and the steam engine;[337]
- → Impacts of previous information technologies: speech and culture, writing, and the printing press; digital services; and communications technologies;[338]
- → Impacts of previous intelligence technologies: price mechanisms in a free market, language, bureaucracy, peer review in science, and evolved institutions like the justice system and law;[339]
- → Impacts of previous labor-substitution technologies as they compare to the possible societal impacts of large language models.[340]

Historical cases of particular dangers or risks from new technologies

Historical precedents for particular types of dangers or threat models from technologies:

---

[332] The distinction between "narrowly transformative," "transformative," and "radically transformative" is found in Gruetzemacher, Ross, and Jess Whittlestone. 'The Transformative Potential of Artificial Intelligence'. *Futures* 135 (2022): 102884. https://doi.org/10.1016/j.futures.2021.102884.

[333] Ding, Jeffrey, and Allan Dafoe. 'Engines of Power: Electricity, AI, and General-Purpose, Military Transformations'. *European Journal of International Security*, 7 February 2023, 1–18. https://doi.org/10.1017/eis.2023.1.

[334] Garfinkel, Ben. 'The Impact of Artificial Intelligence: A Historical Perspective'. In *The Oxford Handbook of AI Governance*, edited by Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, and Baobao Zhang, 0. Oxford University Press, 2023. https://doi.org/10.1093/oxfordhb/9780197579329.013.5.

[335] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press, 2022. https://docs.google.com/document/d/e/2PACX-1vQOQ0EBIaEu_LaJqWvdPKu8xlmrOCM6h6gq7eFHnN0Y2GPYoodQjLeilxQ8SUwnbVThXc0k_jCIsCX1/pub.

[336] The former term is by Garfinkel; the latter by Whittlestone & Gruetzemacher.

[337] Garfinkel, Benjamin. 'The Impact of Artificial Intelligence: A Historical Perspective', 2022. https://docs.google.com/document/d/1I13_0o3kUe1AVQNfevOF9sHpc4mCQkuFDxOXFj_4g-I/edit?.

[338] Mentioned in: Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press, 2022. https://docs.google.com/document/d/e/2PACX-1vQOQ0EBIaEu_LaJqWvdPKu8xlmrOCM6h6gq7eFHnN0Y2GPYoodQjLeilxQ8SUwnbVThXc0k_jCIsCX1/pub.

[339] Ibid.

[340] See informally: dynomight. 'Historical Analogies for Large Language Models'. *DYNOMIGHT INTERNET NEWSLETTER* (blog), 8 December 2022. https://dynomight.substack.com/p/llms.

- → Human-machine interface risks and failures around complex technologies: various "normal accidents" in diverse industries and domains, most notably nuclear power;[341]
- → Technology misuse risks: the proliferation of easily available hacking tools, such as the "Blackshades Remote Access Tool,"[342] but see also the counterexample of non-use of an (apparent) decisive strategic advantage: the brief US nuclear monopoly;[343]
- → Technological "structural risks": the role of technologies in lowering the threshold for war initiation such as the alleged role of railways in inducing swift, all-or-none military mobilization schedules and precipitating escalation to World War I.[344]

Historical cases of value changes as a result of new technologies

Historical precedents for technologically induced value erosion or value shifts:

- → Shared values eroded by pressures of global economic competition: "sustainability, decentralized technological development, privacy, and equality";[345]
- → Technological progress biasing the development of states towards welfare-degrading (inegalitarian and autocratic) forms: agriculture, bronze working, chariots, and cavalry;[346]
- → Technological progress biasing the development of states towards welfare-promoting forms: ironworking, ramming warships, and industrial revolution;[347]
- → Technological progress leading to gradual shifts in societal values: changes in the prevailing technology of energy capture driving changes in societal views on violence, equality, and fairness;[348] demise of dueling and honor culture after (low-skill) pistols replaced (high-skill) swords; changes in

[341] Maas, Matthijs M. 'Regulating for "Normal AI Accidents": Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment'. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 223–28. AIES '18. New York, NY, USA: Association for Computing Machinery, 2018. https://doi.org/10.1145/3278721.3278766.; Williams, Robert, and Roman Yampolskiy. 'Understanding and Avoiding AI Failures: A Practical Guide'. *Philosophies* 6, no. 3 (September 2021): 53. https://doi.org/10.3390/philosophies6030053.; Dietterich, Thomas G. 'Robust Artificial Intelligence and Robust Human Organizations'. *ArXiv:1811.10840 [Cs]*, 27 November 2018. http://arxiv.org/abs/1811.10840.

[342] Hayward, Keith J, and Matthijs M Maas. 'Artificial Intelligence and Crime: A Primer for Criminologists'. *Crime, Media, Culture* 17, no. 2 (30 June 2020): 209–33. https://doi.org/10.1177/1741659020917434. Pg. 10.

[343] Branwen, Gwern. 'Slowing Moore's Law: How It Could Happen', 16 March 2012. https://www.gwern.net/Slowing-Moores-Law. (see subsection "Case-study: Suppressing Nuclear Weapons").

[344] Zwetsloot, Remco, and Allan Dafoe. 'Thinking About Risks From AI: Accidents, Misuse and Structure'. Lawfare, 11 February 2019. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.; drawing on the famous argument in: Van Evera, Stephen. 'The Cult of the Offensive and the Origins of the First World War'. *International Security* 9, no. 1 (1984): 58–107. https://doi.org/10.2307/2538636. However, this interpretation remains contested. See also: Lieber, Keir A. *War and the Engineers: The Primacy of Politics over Technology*. 1 edition. Ithaca; London: Cornell University Press, 2008; Snyder, Jack, and Keir A. Lieber. 'Defensive Realism and the "New" History of World War I'. *International Security* 33, no. 1 (26 June 2008): 174–94. https://doi.org/10.1162/isec.2008.33.1.174.

[345] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press, 2022. https://docs.google.com/document/d/e/2PACX-1vQOQ0EBIaEu_LaJqWvdPKu8xlmrOCM6h6gq7eFHnN0Y2GPYoodQjLeilxQ8SUwnbVThXc0k_jCIsCX1/pub.

[346] MacInnes, Morgan, Ben Garfinkel, and Allan Dafoe. 'Anarchy as Architect: Competitive Pressure, Technology, and the Internal Structure of States', (under review 2023). pg 21.

[347] Ibid.

[348] Danaher, John. 'Axiological Futurism: The Systematic Study of the Future of Values'. *Futures* 132 (1 September 2021): 102780. https://doi.org/10.1016/j.futures.2021.102780. (drawing on: Morris, Ian, *Foragers, Farmers, and Fossil Fuels: How Human Values Evolve*. Edited by Stephen Macedo. Updated ed. edition. Princeton: Princeton University Press, 2015).

sexual morality after the appearance of contraceptive technology; changes in attitudes towards farm animals after the rise of meat replacements; and the rise of the plough as a driver of diverging gender norms.[349]

Historical cases of the disruptive effects on law and governance from new technologies

Historical precedents for effects of new technology on governance tools:

- → Technological changes disrupting or eroding the legal integrity of earlier (treaty) regimes: submarine warfare;[350] implications of cyberwarfare for international humanitarian law;[351] the Soviet Fractional Orbital Bombardment System (FOBS) evading the 1967 Outer Space Treaty's ban on stationing WMDs "in orbit";[352] the mid-2010's US "superfuze" upgrades to its W76 nuclear warheads, massively increasing their counterforce lethality against missile silos without adding a new warhead, missile, or submarine, formally complying with arms control regimes like New START;[353] and various other cases;[354]

- → Technologies strengthening international law: satellites strengthening monitoring with treaty compliance,[355] communications technology strengthening the role of non-state and civil-society actors.[356]

---

[349] See generally: Hopster, J. K. G., C. Arora, C. Blunden, C. Eriksen, L. E. Frank, J. S. Hermann, M. B. O. T. Klenk, E. R. H. O'Neill, and S. Steinert. 'Pistols, Pills, Pork and Ploughs: The Structure of Technomoral Revolutions'. *Inquiry* 0, no. 0 (8 July 2022): 1–33. https://doi.org/10.1080/0020174X.2022.2090434. Hopster, Jeroen K. G., and Matthijs M. Maas. 'The Technology Triad: Disruptive AI, Regulatory Gaps and Value Change'. *AI and Ethics*, 28 June 2023. https://doi.org/10.1007/s43681-023-00305-5.

[350] Crootof, Rebecca. 'Jurisprudential Space Junk: Treaties and New Technologies'. In *Resolving Conflicts in the Law*, edited by Chiara Giorgetti and Natalie Klein, 106–29, 2019. https://brill.com/view/book/edcoll/9789004316539/BP000015.xml.

[351] Eichensehr, Kristen E. 'Cyberwar & International Law Step Zero'. *Texas International Law Journal* 50, no. 2 (2015): 357–80. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2611198

[352] Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. Pg. 203. Drawing on: Garthoff, Raymond L. 'Banning the Bomb in Outer Space'. *International Security* 5, no. 3 (1980): 25–40. https://doi.org/10.2307/2538418. Gyűrösi, Miroslav. 'The Soviet Fractional Orbital Bombardment System Program'. Air Power Australia, 2 January 2010. http://www.ausairpower.net/APA-Sov-FOBS-Program.html.

[353] Kristensen, Hans M., Matthew McKinzie, and Theodore A. Postol. 'How US Nuclear Force Modernization Is Undermining Strategic Stability: The Burst-Height Compensating Super-Fuze'. *Bulletin of the Atomic Scientists* (blog), 1 March 2017. https://thebulletin.org/how-us-nuclear-force-modernization-undermining-strategic-stability-burst-height-compensating-super10578. As discussed in: Maas, Matthijs M. 'Innovation-Proof Governance for Military AI? How I Learned to Stop Worrying and Love the Bot'. *Journal of International Humanitarian Legal Studies* 10, no. 1 (2019): 129–57. https://doi.org/10.1163/18781527-01001006.

[354] Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. Pg. 193-195. See also Crootof, Rebecca. 'Regulating New Weapons Technology'. In *The Impact of Emerging Technologies on the Law of Armed Conflict*, edited by Eric Talbot Jensen and Ronald T.P. Alcala, 1–25. Oxford University Press, 2019. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190915322.001.0001/oso-9780190915322-chapter-1.

[355] Coe, Andrew J., and Jane Vaynman. 'Why Arms Control Is So Rare'. *American Political Science Review* 114, no. 2 (May 2020): 342–55. https://doi.org/10.1017/S000305541900073X.; Vaynman, Jane. 'Better Monitoring and Better Spying: The Implications of Emerging Technology for Arms Control'. *Texas National Security Review* 4, no. 4 (23 September 2021). https://tnsr.org/2021/09/better-monitoring-and-better-spying-the-implications-of-emerging-technology-for-arms-control/.

[356] Maas (ibid. Pg. 224-227). See also: Picker, Colin B. 'A View from 40,000 Feet: International Law and the Invisible Hand of Technology'. *Cardozo Law Review* 23 (2001): 151–219. https://papers.ssrn.com/abstract=987524

Lessons from the history of societal reactions to new technologies
Historical cases that (potentially) provide insights into how societies are likely to perceive, react to, or regulate new technologies.

Historical reactions to and regulations of new technologies

Historical precedents for how key actors are likely to view, treat, or regulate AI:

- → The relative roles of various US actors in shaping the development of past strategic general-purpose technologies: biotech, aerospace tech, and cryptography;[357]
- → Overall US government policy towards perceived "strategic assets": oil[358] and early development of US nuclear power regulation;[359]
- → The historical use of US antitrust law motivated by national security considerations: various cases over the last century;[360]
- → Early regulation of an emerging general-purpose technology: electricity regulation in the US;[361]
- → Previous instances of AI development becoming framed as an "arms race" or competition: 1980's "race" between the US and Japan's Fifth Generation Computer Systems (FGCS) project;[362]
- → Regulation of the "safety" of foundational technology industries, public infrastructures, and sectors: UK regulation of sectors such as medicines and medical devices, food, financial services, transport (aviation & road and rail), energy, and communications;[363]
- → High-level state actors buy-in to ambitious early-stage proposals for world control and development of powerful technology: Initial "Baruch Plan" for world control of nuclear weapons (eventually

[357] Leung, Jade. 'Who Will Govern Artificial Intelligence? Learning from the History of Strategic Politics in Emerging Technologies'. University of Oxford, 2019. https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665.
[358] Ding, Jeffrey, and Allan Dafoe. 'The Logic of Strategic Assets: From Oil to AI'. *Security Studies*, 3 June 2021, 1–31. https://doi.org/10.1080/09636412.2021.1915583.
[359] Walker, J. Samuel, and Thomas R. Wellock. 'A Short History of Nuclear Regulation, 1946–2009'. U.S. Nuclear Regulatory Commission, October 2010. https://www.nrc.gov/docs/ML1029/ML102980443.pdf. I thank Aishwarya Saxena for discussions and suggestions on this topic.
[360] O'Keefe, Cullen. 'How Will National Security Considerations Affect Antitrust Decisions in AI? An Examination of Historical Precedents'. Future of Humanity Institute, 7 July 2020. https://www.fhi.ox.ac.uk/wp-content/uploads/How-Will-National-Security-Considerations-Affect-Antitrust-Decisions-in-AI-Cullen-OKeefe.pdf.
[361] Di Cooke, and Sam Clarke. 'The "Old AI": Lessons for AI Governance from the Early Days of Electricity Regulation'. *Effective Altruism Forum*. 2022. https://forum.effectivealtruism.org/posts/k73qrirnxcKtKZ4ng/the-old-ai-lessons-for-ai-governance-from-early-electricity-1 Drawing among others on: Isser, Steve. *Electricity Restructuring in the United States: Markets and Policy from the 1978 Energy Act to the Present*. Cambridge University Press, 2015.
[362] Garvey, Colin. 'Artificial Intelligence and Japan's Fifth Generation'. Pacific Historical Review 88, no. 4 (1 November 2019): 619–58. https://doi.org/10.1525/phr.2019.88.4.619. Garvey, Shunryu. '"AI for Social Good"; and the First AI Arms Race: Lessons from Japan's Fifth Generation Computer Systems (FGCS) Project'. "AI for Social Good" and the First AI Arms Race: Lessons from Japan's Fifth Generation Computer Systems Project, 1 January 2020. https://www.academia.edu/43348265/_AI_for_Social_Good_and_the_First_AI_Arms_Race_Lessons_from_Japans_Fifth_Generation_Computer_Systems_FGCS_Project.
[363] Smakman, Julia, Matt Davies, and Michael Birtwistle. 'Mission Critical: Lessons from Relevant Sectors for AI Safety'. Ada Lovelace Institute, 31 October 2023. https://www.adalovelaceinstitute.org/policy-briefing/ai-safety/.

failed);[364] extensive early proposals for world control of airplane technology (eventually failed);[365] and repeated (private and public) US offers to the Soviet Union for a joint US-USSR moon mission, including a 1963 UN General Assembly offer by John F. Kennedy to convert the Apollo lunar landing program into a joint US-Soviet moon expedition (initially on-track, with Nikita Khruschev eager to accept the offer; however, Kennedy was assassinated a week after the offer, the Soviets were too suspicious of similar offers by the Johnson administration, and Khruschev was removed from office by coup in 1964);[366]

→ Sustained failure of increasingly more powerful technologies to deliver their anticipated social outcomes: sustained failure of the "Superweapon Peace" idea—the recurring idea that certain weapons of radical destructiveness (nuclear and non-nuclear) may force an end to war by rendering it too destructive to contemplate;[367]

→ Strong public and policy reactions to "warning shots" of a technology being deployed: Sputnik launch and Hiroshima bombing;[368]

→ Strong public and policy reactions to publicly visible accidents involving a new technology: Three Mile Island meltdown,[369] COVID-19 pandemic,[370] and automotive and aviation industries;[371]

→ Regulatory backlash and path dependency: case of genetically modified organism (GMO) regulations in the US vs. the EU;[372]

---

[364] Zaidi, Waqar, and Allan Dafoe. 'International Control of Powerful Technology: Lessons from the Baruch Plan'. Center for the Governance of AI, Future of Humanity Institute, March 2021. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/International-Control-of-Powerful-Technology-Lessons-from-the-Baruch-Plan-Zaidi-Dafoe-2021.pdf.; see generally Baratta, Joseph Preston. *The Politics of World Federation: United Nations, UN Reform, Atomic Control*. Greenwood Publishing Group, 2004.; The analogy is also drawn in Dewey, Daniel. 'Long-Term Strategies for Ending Existential Risk from Fast Takeoff'. In *Risks of Artificial Intelligence*. Chapman and Hall/CRC, 2015. https://www.taylorfrancis.com/chapters/edit/10.1201/b19187-14/long-term-strategies-ending-existential-risk-fast-takeoff-daniel-dewey. Pg. 7-8.

[365] Zaidi, Waqar H. *Technological Internationalism and World Order: Aviation, Atomic Energy, and the Search for International Peace, 1920–1950*. Cambridge ; New York, NY: Cambridge University Press, 2021.

[366] Sietzen, Frank. 'Soviets Planned to Accept JFK's Joint Lunar Mission Offer'. SpaceCast News Service, 2 October 1997. https://www.spacedaily.com/news/russia-97h.html.; as also discussed by Richard Ngo in: Bensinger, Rob. 'Ngo's View on Alignment Difficulty'. Machine Intelligence Research Institute, 15 December 2021. https://intelligence.org/2021/12/14/ngos-view-on-alignment-difficulty/.

[367] Renic, Neil C. 'Superweapons and the Myth of Technological Peace'. *European Journal of International Relations*, 15 November 2022, 13540661221136764. https://doi.org/10.1177/13540661221136764.

[368] See influentially Boyer, Paul. *By the Bomb's Early Light*. The University of North Carolina Press, 1985. https://uncpress.org/book/9780807844809/by-the-bombs-early-light/. Lente, Dick van, ed. *The Nuclear Age in Popular Media*. New York: Palgrave Macmillan US, 2012. https://doi.org/10.1057/9781137086181. I thank Lara Thurnherr for these suggestions.

[369] Gabs, Nick. 'Lessons from Three Mile Island for AI Warning Shots'. EA Forum, 26 September 2022. https://forum.effectivealtruism.org/posts/NyCHoZGGw5YssvDJB/lessons-from-three-mile-island-for-ai-warning-shots.

[370] Krakovna, Victoria. 'Possible Takeaways from the Coronavirus Pandemic for Slow AI Takeoff'. *AI Alignment Forum* (blog), 31 May 2020. https://www.alignmentforum.org/posts/wTKjRFeSjKLDSWyww/possible-takeaways-from-the-coronavirus-pandemic-for-slow-ai.; Soares, Nate. 'Warning Shots Probably Wouldn't Change The Picture Much'. Alignment Forum, 6 October 2022. https://www.alignmentforum.org/posts/idipkijjz5PoxAwju/warning-shots-probably-wouldn-t-change-the-picture-much. See generally: Liu, Hin-Yan, Kristian Lauta, and Matthijs Maas. 'Apocalypse Now?: Initial Lessons from the Covid-19 Pandemic for the Governance of Existential and Global Catastrophic Risks'. *Journal of International Humanitarian Legal Studies* 11, no. 2 (13 August 2020): 295–310. https://doi.org/10.1163/18781527-01102004.

[371] Lupo, Giampiero. 'Risky Artificial Intelligence: The Role of Incidents in the Path to AI Regulation'. *Law, Technology and Humans* 5, no. 1 (30 May 2023): 133–52. https://doi.org/10.5204/lthj.2682.

[372] Grotto, Andrew. 'Genetically Modified Organisms: A Precautionary Tale For AI Governance'. *AI Pulse*, 24 January 2019. https://aipulse.org/genetically-modified-organisms-a-precautionary-tale-for-ai-governance-2/.

→ "Regulatory capture" and/or influence of industry actors on tech policy, the role of the US military industrial complex in perpetuating the "bomber gap" and "missile gap" myths,[373] and undue corporate influence in the World Health Organisation during the 2009 H1N1 pandemic;[374]

→ State norm "antipreneurship" (actions aiming to preserve the prevailing global normative status quo at the global level against proposals for new regulation or norm-setting): US resistance to proposed global restraints on space weapons, between 2000 and the present, utilizing a range of diplomatic strategies and tactics to preserve a permissive international legal framework governing outer space.[375]

Lessons from the history of attempts to initiate technology governance

Historical cases that (potentially) provide insights into when efforts to initiate governance intervention on emerging technologies are likely to be successful and into the efficacy of various pathways towards influencing key actors to deploy regulatory levers in response.

Historical failures to initiate or shape technology governance

Historical cases where a fear of false positives slowed (plausibly warranted) regulatory attention or intervention:

→ Failure to act in spite of growing evidence: a review of nearly 100 cases of environmental issues where the precautionary principle was raised, concluding that fear of false positives has often stalled action even though (i) false positives are rare and (ii) there was enough evidence to suggest that a lack of regulation could lead to harm.[376]

Historical cases of excessive hype leading to (possibly) premature regulatory attention or intervention:

→ Premature (and possibly counterproductive) legal focus on technologies that eventually took much longer to develop than anticipated: Weather modification technology,[377] deep seabed mining,[378]

[373] Kemp, Luke. 'Agents of Doom: Who Is Creating the Apocalypse and Why'. *BBC Future*, 26 October 2021. https://www.bbc.com/future/article/20211014-agents-of-doom-who-is-hastening-the-apocalypse-and-why.;

[374] See generally; Deshman, Abigail C. 'Horizontal Review between International Organizations: Why, How, and Who Cares about Corporate Regulatory Capture'. *European Journal of International Law* 22, no. 4 (1 November 2011): 1089–1113. https://doi.org/10.1093/ejil/chr093. As discussed in: Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890.

[375] See generally: Bower, Adam, and Jeffrey S. Lantis. 'Contesting the Heavens: US Antipreneurship and the Regulation of Space Weapons'. *European Journal of International Security*, 8 February 2023, 1–22. https://doi.org/10.1017/eis.2023.2.

[376] European Environment Agency. *Late Lessons from Early Warnings :Science, Precaution, Innovation.* LU: Publications Office, 2013. https://data.europa.eu/doi/10.2800/73322.; see also Harremoës, Poul, ed. *Late Lessons from Early Warnings: The Precautionary Principle, 1896-2000*. Environmental Issue Report, no. 22. Copenhagen, Denmark: European Environment Agency, 2001. I thank José Jaime Villalobos and Andrew Mazibrada for insights on this case.

[377] See generally: Fleming, James Rodger. 'The Pathological History of Weather and Climate Modification: Three Cycles of Promise and Hype'. *Historical Studies in the Physical and Biological Sciences* 37, no. 1 (1 September 2006): 3–25. https://doi.org/10.1525/hsps.2006.37.1.3.

[378] Picker, Colin B. 'A View from 40,000 Feet: International Law and the Invisible Hand of Technology'. *Cardozo Law Review* 23 (2001): 151–219. https://papers.ssrn.com/abstract=987524

self-driving cars,[379] virtual and augmented reality,[380] and other technologies charted under the Gartner Hype Cycle reports.[381]

Historical successes for pathways in shaping technology governance

Historical precedents for successful action towards understanding and responding to the risks of emerging technologies, influencing key actors to deploy regulatory levers:

- → Relative success in long-range technology forecasting: some types of forecasts for military technology that achieved reasonable accuracy decades out;[382]
- → Success in anticipatory governance: history of "prescient actions" in urging early action against risky new technologies, such as Leo Szilard's warning of the dangers of nuclear weapons[383] and Alexander Fleming's 1945 warning of the risk of antibiotic resistance;[384]
- → Successful early action to set policy for safe innovation in a new area of science:[385] the 1967 Outer Space Treaty, UK's Warnock Committee and Human Embryology Act 1990, the Internet Corporation for Assigned Names and Numbers (ICANN);

[379] See for example Chafkin, Max. 'Even After $100 Billion, Self-Driving Cars Are Going Nowhere'. *Bloomberg.Com*, 6 October 2022. https://www.bloomberg.com/news/features/2022-10-06/even-after-100-billion-self-driving-cars-are-going-nowhere.; but see also: Templeton, Brad. 'Reports Of The Death Of Self-Driving Cars Are Greatly Exaggerated'. *Forbes*, 15 November 2022, sec. Transportation. https://www.forbes.com/sites/bradtempleton/2022/11/15/reports-of-the-death-of-self-driving-cars-are-greatly-exaggerated/.

[380] Liao, Tony, and Andrew Iliadis. 'A Future so Close: Mapping 10 Years of Promises and Futures across the Augmented Reality Development Cycle'. *New Media & Society* 23, no. 2 (1 February 2021): 258–83. https://doi.org/10.1177/1461444820924623.

[381] E.g., see: Gartner. 'What's New in the 2022 Gartner Hype Cycle for Emerging Technologies'. Gartner, 2023. https://www.gartner.com/en/articles/what-s-new-in-the-2022-gartner-hype-cycle-for-emerging-technologies.; though for a critique, see also: Steinert, Martin, and Larry Leifer. 'Scrutinizing Gartner's Hype Cycle Approach'. In *PICMET 2010 Proceedings*. Phuket, Thailand, 2010. https://www.researchgate.net/profile/Martin_Steinert/publication/224182916_Scrutinizing_Gartner%27s_hype_cycle_approach/links/543005400cf29bbc1273c7e1/Scrutinizing-Gartners-hype-cycle-approach.pdf.

[382] Kott, Alexander, and Philip Perconti. 'Long-Term Forecasts of Military Technologies for a 20-30 Year Horizon: An Empirical Assessment of Accuracy'. *ArXiv:1807.08339 [Cs]*, 22 July 2018. http://arxiv.org/abs/1807.08339. But for discussion of the methodological differences in evaluating these (and other) historical long-range forecasting exercises, see also: Muelhauser, Luke. 'Evaluation of Some Technology Forecasts from "The Year 2000"'. *Open Philanthropy* (blog), July 2017. https://www.openphilanthropy.org/research/evaluation-of-some-technology-forecasts-from-the-year-2000/. And Muelhauser, Luke. 'How Feasible Is Long-Range Forecasting?' *Open Philanthropy* (blog), 10 October 2019. https://www.openphilanthropy.org/research/how-feasible-is-long-range-forecasting/.

[383] Grace, Katja. 'Leó Szilárd and the Danger of Nuclear Weapons: A Case Study in Risk Mitigation'. Technical Report. Berkeley, CA: Machine Intelligence Research Institute, October 2015. https://intelligence.org/files/SzilardNuclearWeapons.pdf.

[384] Korzekwa, Rick. 'Preliminary Survey of Prescient Actions'. AI Impacts, 3 April 2020. https://aiimpacts.org/survey-of-prescient-actions/. (briefly surveying 20 possible cases).

[385] See also Harding, Verity. 'Lessons from History: What Can Past Technological Breakthroughs Teach the AI Community Today', 2020. https://www.bennettinstitute.cam.ac.uk/blog/lessons-history-what-can-past-technological-breakt/.

→ Governmental reactions and responses to new risks as they emerge: the 1973 Oil Crisis, the 1929–1933 Great Depression,[386] the 2007–2009 financial crisis,[387] the COVID-19 pandemic;[388]

→ How effectively other global risks motivated action in response, and how cultural and intellectual orientations influence perceptions: biotechnology, nuclear weapons, global warming, and asteroid collision;[389]

→ The impact of cultural media (film, etc.) on priming policymakers to risks:[390] the role of *The Day After* in motivating Cold War efforts towards nuclear arms control,[391] of the movies *Deep Impact* and *Armageddon* in shaping perceptions of the importance of asteroid defense,[392] of the novel *Ghost Fleet* in shaping Pentagon perceptions of the importance of emerging technologies to war,[393] of *Contagion* in priming early UK policy responses to COVID-19,[394] of *Mission Impossible: Dead Reckoning: Part One* in deepening President Biden's concerns over AI prior to signing a landmark 2023 Executive Order.[395]

→ The impact of different analogies or metaphors in framing technology policy:[396] for example, the US military's emphasis on framing the internet as "cyberspace" (i.e., just another "domain" of conflict)

[386] Eigner, Peter, and Thomas S. Umlauft. 'The Great Depression(s) of 1929-1933 and 2007-2009? Parallels, Differences and Policy Lessons'. Hungarian Academy of Science MTA-ELTE Crisis History Working Paper No. 2. Rochester, NY, 1 July 2015. https://doi.org/10.2139/ssrn.2612243. I thank Lara Thurnherr for these and other suggestions.

[387] Muehlhauser, Luke. 'How Well Will Policy-Makers Handle AGI? (Initial Findings)'. Machine Intelligence Research Institute, 12 September 2013. https://intelligence.org/2013/09/12/how-well-will-policy-makers-handle-agi-initial-findings/. (also discussing a set of other examples or analogies, which are however dismissed for not being sufficiently similar to AGI risk on various grounds).

[388] Krakovna, Victoria. 'Possible Takeaways from the Coronavirus Pandemic for Slow AI Takeoff'. *AI Alignment Forum* (blog), 31 May 2020. https://www.alignmentforum.org/posts/wTKjRFeSjKLDSWyww/possible-takeaways-from-the-coronavirus-pandemic-for-slow-ai.; Soares, Nate. 'Warning Shots Probably Wouldn't Change The Picture Much'. Alignment Forum, 6 October 2022. https://www.alignmentforum.org/posts/idipkijjz5PoxAwju/warning-shots-probably-wouldn-t-change-the-picture-much.

[389] Baum, Seth. 'Lessons for Artificial Intelligence from Other Global Risks'. In *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, 20, 2019.

[390] I thank Oliver Guest for suggesting this category.

[391] Gaulkin, Thomas. 'Facing Nuclear Reality, 35 Years after The Day After'. *Bulletin of the Atomic Scientists* (blog), 13 December 2018. https://thebulletin.org/2018/12/facing-nuclear-reality-35-years-after-the-day-after/.; Knoblauch, William M. *Nuclear Freeze in a Cold War: The Reagan Administration, Cultural Activism, and the End of the Arms Race*. University of Massachusetts Press, 2017. https://doi.org/10.2307/j.ctv346v1z.; see also Feldman, Stanley, and Lee Sigelman. 'The Political Impact of Prime-Time Television: "The Day After"'. *The Journal of Politics* 47, no. 2 (1985): 556–78. https://doi.org/10.2307/2130896.

[392] Wiblin, Robert, and Keiran Harris. 'Carl Shulman on the Common-Sense Case for Existential Risk Work and Its Practical Implications'. 80,000 Hours Podcast. Accessed 11 October 2021. https://80000hours.org/podcast/episodes/carl-shulman-common-sense-case-existential-risks/.

[393] Luce, Dan De. 'A Novel About War With China Strikes a Chord at the Pentagon'. *Foreign Policy* (blog), 15 May 2016. https://foreignpolicy.com/2016/05/15/a-novel-about-war-with-china-strikes-a-chord-at-the-pentagon/.; Harper, Jon. 'Pentagon Betting on New Technologies to Foil Future Adversaries'. *National Defense* 101, no. 756 (2016): 26–29. https://www.jstor.org/stable/27021585

[394] Forrest, Adam. 'Matt Hancock Admits Hollywood Film Contagion Shaped Vaccine Response'. The Independent, 4 February 2021. https://www.independent.co.uk/news/uk/politics/covid-vaccine-strategy-hancock-contagion-movie-b1796923.html.

[395] Hagy, Paige, and Rachyl Jones. 'The White House Just Revealed a Key Factor Driving Biden's New Order to Rein in AI: The Latest Tom Cruise "Mission: Impossible" Movie'. Fortune, 1 November 2023. https://fortune.com/2023/11/01/biden-ai-executive-order-tom-cruise-mission-impossible-movie/.

[396] Maas, Matthijs, 'AI is like… A literature review of AI metaphors and why they matter for policy.' Institute for Law & AI. AI Foundations Report 2. (October 2023). https://www.legalpriorities.org/research/ai-policy-metaphors (discussing this case, and various other cases relating to both internet policy and the regulation of AI).

led to strong consequences institutionally (supporting the creation of the US Cyber Command) as well as for how international law has subsequently been applied to cyber operations;[397]

→ The role of "epistemic communities" of experts in advocating for international regulation or agreements,[398] specifically their role in facilitating nonproliferation treaties and arms control agreements for nuclear weapons[399] and anti-ballistic missile systems,[400] as well as the history of the earlier era of arms control agreements;[401]

→ Attempted efforts towards international control of new technology: early momentum but ultimate failure of the Baruch Plan for world control of nuclear weapons[402] and the failure of world control of aviation in 1920s;[403]

→ Policy responses to past scientific breakthroughs, and the role of geopolitics vs. expert engagement: the 1967 UN Outer Space Treaty, the UK's Warnock Committee and the Human Fertilisation and Embryology Act 1990, the establishment of the Internet Corporation for Assigned Names and Numbers (ICANN), and the European ban on GMO crops;[404]

→ The role of activism and protests in spurring nonproliferation and moratoria in spurring nuclear nonproliferation agreements and nuclear test bans;[405] the role of activism (in response to "trigger events") in achieving a de facto moratorium on genetically modified crops in Europe in the late

[397] Branch, Jordan. 'What's in a Name? Metaphors and Cybersecurity'. *International Organization* 75 (2021): 39–70. https://doi.org/10.1017/S002081832000051X.

[398] Though for a discussion of cases where epistemic communities failed, see also: Cross, Mai'a K. Davis. 'The Limits of Epistemic Communities: EU Security Agencies'. *Politics and Governance* 3, no. 1 (31 March 2015): 90. https://doi.org/10.17645/pag.v3i1.78. (discussing the surprisingly limited influence of the European Defence Agency (EDA) and EU Intelligence Analysis Centre (IntCen) at shaping EU security policy).

[399] See Kutchesfahani, Sara Z. *Politics and the Bomb: The Role of Experts in the Creation of Cooperative Nuclear Non-Proliferation Agreements*. New York: Routledge, 2013. https://doi.org/10.4324/9780203116500. And previously Kutchesfahani, Sara Zahra. 'Politics & The Bomb: Exploring the Role of Epistemic Communities in Nuclear Non-Proliferation Outcomes.' UCL (University College London), 2010. https://core.ac.uk/download/pdf/1862576.pdf. I thank Charlie Harrison for suggestions here. And for work on the efforts by scientists during the early nuclear age to advocate for (ultimately unsuccessful) proposals for global control of nuclear weapons, see: Zaidi, Waqar, and Allan Dafoe. 'International Control of Powerful Technology: Lessons from the Baruch Plan'. Center for the Governance of AI, Future of Humanity Institute, March 2021. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/International-Control-of-Powerful-Technology-Lessons-from-the-Baruch-Plan-Zaidi-Dafoe-2021.pdf. Ord, Toby. 'Lessons from the Development of the Atomic Bomb'. Center for the Governance of AI, November 2022. https://www.governance.ai/research-paper/lessons-atomic-bomb-ord.

[400] Maas, Matthijs M. 'How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons'. *Contemporary Security Policy* 40, no. 3 (6 February 2019): 285–311. https://doi.org/10.1080/13523260.2019.1576464. Drawing significantly on Adler, Emanuel. 'The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control'. *International Organization* 46, no. 1 (1992): 101–45. https://doi.org/10.1017/S0020818300001466.

[401] Scharre, Paul, and Megan Lamberth. 'Artificial Intelligence and Arms Control'. Center for a New American Security, 12 October 2022. https://www.cnas.org/publications/reports/artificial-intelligence-and-arms-control. (discussing how in the wake of the 1868 St. Petersburg Declaration "states engaged in a flurry of arms control activity, both in the run-up to World War I and in the interwar period before World War II", and deriving lessons for AI arms control).

[402] Zaidi, Waqar, and Allan Dafoe. 'International Control of Powerful Technology: Lessons from the Baruch Plan'. Center for the Governance of AI, Future of Humanity Institute, March 2021. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/International-Control-of-Powerful-Technology-Lessons-from-the-Baruch-Plan-Zaidi-Dafoe-2021.pdf.

[403] Zaidi, Waqar H. *Technological Internationalism and World Order: Aviation, Atomic Energy, and the Search for International Peace, 1920–1950*. Cambridge ; New York, NY: Cambridge University Press, 2021.

[404] Harding, Verity. 'Lessons from History: What Can Past Technological Breakthroughs Teach the AI Community Today', 2020. https://www.bennettinstitute.cam.ac.uk/blog/lessons-history-what-can-past-technological-breakt/.

[405] See Wittner, Lawrence S. 'The Nuclear Freeze and Its Impact'. Arms Control Association, 2010. https://www.armscontrol.org/act/2010_12/LookingBack.; and also: Cooke, Di. 'The Role of Activism in Nuclear Arms Control', 2020. (private draft). I thank Charlie Harrison for suggestions in this area of research.

1990s;[406] in addition, the likely role of protests and public pressure in contributing to abandonment or slowing of various technologies from geoengineering experiments to nuclear weapons, CFCs, and nuclear power;[407]

→ The role of philanthropy and scientists in fostering Track-II diplomacy initiatives: the Pugwash conferences.[408]

Lessons from the historical efficacy of different governance levers

Historical cases that (potentially) provide insights into when different societal (legal, regulatory, and governance) levers have proven effective in shaping technology development and use in desired directions.

Historical failures of technology governance levers

Historical precedents for failed or unsuccessful use of various (domestic and/or international) governance levers for shaping technology:

→ Mixed-success use of soft-law governance tools for shaping emerging technologies: National Telecommunications and Information Administration discussions on mobile app transparency, drone privacy, facial recognition, YourAdChoices, UNESCO declaration on genetics and bioethics, Environmental Management System (ISO 14001), Sustainable Forestry Practices by the Sustainable Forestry Initiative and Forest Stewardship Council, and Leadership in Energy and Environmental Design.[409]

→ Failed use of soft-law governance tools for shaping emerging technologies: Children's Online Privacy Protection Rule, Internet Content Rating Association, Platform for Internet Content Selection, Platform for Privacy Preferences, Do Not Track system, and nanotechnology voluntary data call-in by Australia, the US, and the UK;[410]

→ Failures of narrowly technology-focused approaches to safety engineering: failure of narrow technology-focused approaches to the design of safe cars and in the design and calibration of pulse oximeters during the COVID pandemic, which were mismatched to—and therefore led to dangerous outcomes for—female drivers and darker-skinned patients, respectively, highlighting the role of incorporating human, psychological, and other disciplines;[411]

→ Failures of information control mechanisms at preventing proliferation: selling of nuclear secrets by A.Q. Khan network,[412] limited efficacy of Cold War nuclear secrecy regimes at meaningfully

[406] Harrison, Charlie. 'Go Mobilize? Lessons from GM Protests for Pausing AI'. *EA Forum* (blog), 24 October 2023. https://forum.effectivealtruism.org/posts/6jxrzk99eEjsBxoMA/go-mobilize-lessons-from-gm-protests-for-pausing-ai.
[407] Harrison, Charlie. 'Efficacy of AI Activism: Have We Ever Said No?', Forthcoming 2023.
[408] Martin, Rani. 'The Pugwash Conferences and the Anti-Ballistic Missile Treaty as a Case Study of Track II Diplomacy'. EA Forum, 16 September 2022. https://forum.effectivealtruism.org/posts/ggiCDnYcSKLxwFbBv/the-pugwash-conferences-and-the-anti-ballistic-missile.
[409] Gutierrez, Carlos Ignacio, Gary E. Marchant, and Lucille Tournas. 'Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance'. *JURIMETRICS* 61, no. 1 (29 December 2020). https://doi.org/10.2139/ssrn.3775271.
[410] Ibid.
[411] Vallor, Shannon, and Ewa Luger. 'A Shrinking Path to Safety: How a Narrowly Technical Approach to Align AI with the Public Good Could Fail'. *BRAID UK* (blog), 13 October 2023. https://braiduk.org/a-shrinking-path-to-safety-how-a-narrowly-technical-approach-to-align-ai-with-the-public-good-could-fail.
[412] Laufer, Michael. 'A. Q. Khan Nuclear Chronology'. Carnegie Endowment for International Peace, 7 September 2005. https://carnegieendowment.org/2005/09/07/a.-q.-khan-nuclear-chronology-pub-17420. ; MacCalman, Molly. 'A.Q. Khan Nuclear Smuggling Network'. Journal of Strategic Security 9, no. 1 (March 2016): 104–18. https://doi.org/10.5038/1944-0472.9.1.1506.

constraining proliferation of nuclear weapons,[413] track record of major leaks and hacks of digital information, 2005–present;[414]

→ Failure to transfer (technological) safety techniques, even to allies: in the late 2000s, the US sought to help provide security assistance to Pakistan to help safeguard the Pakistani nuclear arsenal but was unable to transfer permissive action link (PAL) technologies because of domestic legal barriers that forbade export to states that were not part of the Nuclear Non-Proliferation Treaty (NPT);[415]

→ Degradation of previously established export control regimes: Cold War-era US high performance computing export controls struggled to be updated sufficiently quickly to keep pace with hardware advancements.[416] The US initially treated cryptography as a weapon under export control laws, meaning that encryption systems could not be exported for commercial purposes even to close allies and trading partners; however, by the late 1990s, several influences—including the rise of open-source software and European indignation at US spying on their communications—led to new regulations that allowed cryptography to be exported with minimal government interference;[417]

→ "Missed opportunities" for early action against anticipated risks: mid-2000s effort to put "killer robots" on humanitarian disarmament issue agenda, which failed as these were seen as "too speculative";[418]

→ Mixed success of scientific and industry self-regulation: the Asilomar Conference, the Second International Conference on Synthetic Biology, and 2004–2007 failed efforts to develop guidelines for nanoparticles;[419]

→ Sustained failure to establish treaty regimes: various examples, including the international community spending nearly 20 years since 2004 negotiating a new treaty for Biodiversity Beyond National Jurisdiction;[420]

→ Unproductive locking-in of insufficient, "empty" institutions, "face-saving" institutions, or gridlocked mechanisms: history of states creating suboptimal, ill-designed institutions—such as the United Nations Forum on Forests, the Copenhagen Accord on Climate Change, the UN Commission on

---

[413] Wellerstein, Alex. Restricted Data: The History of Nuclear Secrecy in the United States. Chicago, IL: University of Chicago Press, 2021. https://press.uchicago.edu/ucp/books/book/chicago/R/bo15220099.html.

[414] See the compilation in: Muelhauser, Luke. 'Example High-Stakes Information Security Breaches [Public]', June 2020. https://docs.google.com/document/d/1_smEDPWDVIaLuZ14Cm7KLHcWx4LkJ0DCTk8bcHjYy_Y/edit#heading=h.hqf76e8phc7g.

[415] Lewis, Jeffrey. 'No PALs For Paks'. Arms Control Wonk, 2007. https://www.armscontrolwonk.com/archive/201709/no-pals-for-paks/. Khan, Feroz Hassan. 'Nuclear Security in Pakistan: Separating Myth From Reality'. Arms Control Association, 2009. https://www.armscontrol.org/act/2009-07/features/nuclear-security-pakistan-separating-myth-reality.

[416] Picker, Colin B. 'A View from 40,000 Feet: International Law and the Invisible Hand of Technology'. *Cardozo Law Review* 23 (2001): 151–219. https://papers.ssrn.com/abstract=987524

[417] Diffie, Whitfield, and Susan Landau. 'The Export of Cryptography in the 20th and the 21st Centuries'. In The History of Information Security, edited by Karl De Leeuw and Jan Bergstra, 725–36. Amsterdam: Elsevier Science B.V., 2007. https://doi.org/10.1016/B978-044451608-4/50027-4.

[418] Carpenter, Charli. *'Lost' Causes, Agenda Vetting in Global Issue Networks and the Shaping of Human Security*. Ithaca: Cornell University Press, 2014. https://doi.org/10.7591/9780801470363.

[419] jia. 'Case Studies of Self-Governance to Reduce Technology Risk'. EA Forum, 6 April 2021. https://forum.effectivealtruism.org/posts/Xf6QE6txgvfCGvZpk/case-studies-of-self-governance-to-reduce-technology-risk.

[420] Tiller, Rachel, Elizabeth Mendenhall, Elizabeth De Santo, and Elizabeth Nyman. 'Shake It Off: Negotiations Suspended, but Hope Simmering, after a Lack of Consensus at the Fifth Intergovernmental Conference on Biodiversity beyond National Jurisdiction'. *Marine Policy* 148 (1 February 2023): 105457. https://doi.org/10.1016/j.marpol.2022.105457.

Sustainable Development, and the 1980 UN Convention on Certain Conventional Weapons—with mandates that may deprive them of much capacity for policy formulation or implementation;[421]

→ Drawn-out contestation of hierarchical and unequal global technology governance regimes: the Nuclear Non-Proliferation Treaty regime has seen cycles of contestation and challenge by other states;[422]

→ Failures of non-inclusive club governance approaches to nonproliferation: the Nuclear Security Summits (NSS) (2012, 2014, 2016) centered on high-level debates over the stocktaking and securing of nuclear materials. These events saw a constrained list of invited states; as a result, the NSS process was derailed because procedural questions over who was invited or excluded came to dominate discussions (especially at the 2016 Vienna summit), politicizing what had been a technical topic and hampering the extension and take-up of follow-on initiatives by other states.[423]

Historical successes of technology governance levers

Historical precedents for successful use of various governance levers at shaping technology:

→ Effective scientific secrecy around early development of powerful new technologies: early development of the atomic bomb[424] and early computers (Colossus and ENIAC).[425]

→ Successes in the oversight of various safety-critical technologies: track record of "High Reliability Organisations"[426] in addressing emerging risks after initial incidents to achieve very low rates of

[421] On "empty" institutions (discussing these environmental regimes), see: Dimitrov, Radoslav S. 'Empty Institutions in Global Environmental Politics'. *International Studies Review* 22, no. 3 (1 September 2020): 626–50. https://doi.org/10.1093/isr/viz029. On "face-saving" institutions (discussing the CCW), see: Mantilla, Giovanni. 'Deflective Cooperation: Social Pressure and Forum Management in Cold War Conventional Arms Control'. *International Organization* 77, no. 3 (March 2023): 564–98. https://doi.org/10.1017/S0020818322000364.

[422] Egeland, K. 'The Road to Prohibition: Nuclear Hierarchy and Disarmament, 1968–2017'. Http://purl.org/dc/dcmitype/Text, University of Oxford, 2017. https://ora.ox.ac.uk/objects/uuid:b03d68ab-4748-4de7-a2e9-15616de6a05c. Though for a discussion of how global arms control institutions have gradually evolved in ways that have replaced or supplemented old forms of institutional inequality, see Fehl, Caroline. 'Unequal Power and the Institutional Design of Global Governance: The Case of Arms Control'. *Review of International Studies* 40, no. 3 (July 2014): 505–31. https://doi.org/10.1017/S026021051300034X.

[423] Stover, Dawn. 'The Controversial Legacy of the Nuclear Security Summit'. Bulletin of the Atomic Scientists (blog), 4 October 2018. https://thebulletin.org/2018/10/the-controversial-legacy-of-the-nuclear-security-summit/.

[424] Grace, Katja. 'Leó Szilárd and the Danger of Nuclear Weapons: A Case Study in Risk Mitigation'. Technical Report. Berkeley, CA: Machine Intelligence Research Institute, October 2015. https://intelligence.org/files/SzilardNuclearWeapons.pdf.

[425] Napper, Brian. 'Early Computers (1946-51)'. Computer50, 20 August 1999. http://curation.cs.manchester.ac.uk/computer50/www.computer50.org/mark1/contemporary.html. I thank Lara Thurnherr for this suggestion.

[426] Roberts, Karlene H. 'New Challenges in Organizational Research: High Reliability Organizations'. Industrial Crisis Quarterly 3, no. 2 (1 June 1989): 111–25. https://doi.org/10.1177/108602668900300202.; Lekka, Chrysanthi. 'High Reliability Organisations - A Review of the Literature'. Health and Safety Executive, 2011. https://www.hse.gov.uk/research/rrpdf/rr899.pdf. For applications to AI, see: Dietterich, Thomas G. 'Robust Artificial Intelligence and Robust Human Organizations'. arXiv, 27 November 2018. https://doi.org/10.48550/arXiv.1811.10840.; Shneiderman, Ben. 'Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems'. ACM Transactions on Interactive Intelligent Systems 10, no. 4 (16 October 2020): 26:1-26:31. https://doi.org/10.1145/3419764. Critch, Andrew, and David Krueger. 'AI Research Considerations for Human Existential Safety (ARCHES)', 29 May 2020. http://acritch.com/arches/. (pg. 83-84).

errors, such as in air traffic control systems, naval aircraft carrier operations,[427] the aerospace sector, construction, and oil refineries;[428]

→ Successful development of "defense in depth"[429] interventions to lower the risks of accident in specific industries: safe operation of nuclear reactors, chemical plants, aviation, space vehicles, cybersecurity and information security, software development, laboratories studying dangerous pathogens, improvised explosive devices, homeland security, hospital security, port security, physical security in general, control system safety in general, mining safety, oil rig safety, surgical safety, fire management, and health care delivery,[430] and lessons from defense-in-depth frameworks developed in cybersecurity for frontier AI risks;[431]

→ Successful safety "races to the top" in selected industries: Improvements in aircraft safety in the aviation sector;[432]

→ Successful use of risk assessment techniques in safety-critical industries: examination of popular risk identification techniques (scenario analysis, fishbone method, and risk typologies and taxonomies), risk analysis techniques (causal mapping, Delphi technique, cross-impact analysis, bow tie analysis, and system-theoretic process analysis), and risk evaluation techniques (checklists and risk matrices) used in established industries like finance, aviation, nuclear, and biolabs, and how these might be applied in advanced AI companies;[433]

→ Susceptibility of different types of digital technologies to (global) regulation: relative successes and failures of global regulation of different digital technologies that are (1) centralized and clearly material (e.g., submarine cables), (2) decentralized and clearly material (e.g., smart speakers); (3)

[427] Rochlin, Gene I, Todd R La Porte, and Karlene H Roberts. 'The Self-Designing High-Reliability Organization: Aircraft Carrier Flight Operations at Sea' 40, no. 4 (1987): 18. https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=4373&context=nwc-review ; Roberts, Karlene H., Suzanne K. Stout, and Jennifer J. Halpern. 'Decision Dynamics in Two High Reliability Military Organizations'. *Management Science* 40, no. 5 (May 1994): 614–24. https://doi.org/10.1287/mnsc.40.5.614.

[428] Lekka, Chrysanthi, and Caroline Sugden. 'The Successes and Challenges of Implementing High Reliability Principles: A Case Study of a UK Oil Refinery'. *Process Safety and Environmental Protection*, Special Issue: Centenary of the Health and Safety Issue, 89, no. 6 (1 November 2011): 443–51. https://doi.org/10.1016/j.psep.2011.07.003.

[429] For work on a "defense in depth" approach to existential risks generally, see Cotton‐Barratt, Owen, Max Daniel, and Anders Sandberg. 'Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter'. *Global Policy* 11, no. 3 (2020): 271–82. https://doi.org/10.1111/1758-5899.12786.

[430] Listed in: Muelhauser, Luke. 'A Personal Take on Longtermist AI Governance'. EA Forum, 16 July 2021. https://forum.effectivealtruism.org/posts/M2SBwctwC6vBqAmZW/a-personal-take-on-longtermist-ai-governance. (footnote 19).

[431] Ee, Shaun, Joe O'Brien, Zoe Williams, Amanda El-Dakhakhni, Michael Aird, and Alex Lintz. 'Adapting Cybersecurity Frameworks to Manage Frontier AI Risks: A Defense-In-Depth Approach'. Institute for AI Policy and Strategy (IAPS), 13 October 2023. https://www.iaps.ai/research/adapting-cybersecurity-frameworks.

[432] Hunt, Will. 'The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry'. Center for Long-Term Cybersecurity, August 2020. https://cltc.berkeley.edu/publication/new-report-the-flight-to-safety-critical-ai-lessons-in-ai-safety-from-the-aviation-industry/.

[433] Koessler, Leonie, and Jonas Schuett. 'Risk Assessment at AGI Companies: A Review of Popular Risk Assessment Techniques from Other Safety-Critical Industries'. arXiv, 17 July 2023. http://arxiv.org/abs/2307.08823.

centralized and seemingly immaterial (e.g., search engines), and (4) decentralized and seemingly immaterial (e.g., Bitcoin protocol);[434]

→ Use of confidence-building measures to stabilize relations and expectations: 1972 Incidents at Sea Agreement[435] and the 12th–19th century development of Maritime Prize Law;[436]

→ Successful transfer of developed safety techniques, even to adversaries: the US "leaked" PAL locks on nuclear weapons to the Soviet Union;[437]

→ Effective nonproliferation regimes: for nuclear weapons, a mix of norms, treaties, US "strategies of inhibition,"[438] supply-side export controls,[439] and domestic politics factors[440] have produced an imperfect but remarkably robust track record of nonproliferation.[441] Indeed, based on IAEA databases there have historically been 74 states that decided to build or use nuclear reactors, of which 69 have at some time been considered potentially able to pursue nuclear weapons, and of which 10 went nuclear and 7 ran but abandoned a program, and for 14–23, evidence exists of a considered decision not to use their infrastructure to pursue nuclear weapons;[442]

---

[434] See generally: Beaumier, Guillaume, Kevin Kalomeni, Malcolm Campbell‐Verduyn, Marc Lenglet, Serena Natile, Marielle Papin, Daivi Rodima‐Taylor, Arthur Silve, and Falin Zhang. 'Global Regulations for a Digital Economy: Between New and Old Challenges'. *Global Policy* 11, no. 4 (September 2020): 515–22. https://doi.org/10.1111/1758-5899.12823.; see also the analysis of the regulatability of hardware in Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims'. *ArXiv:2004.07213 [Cs]*, 15 April 2020. http://arxiv.org/abs/2004.07213.

[435] Ruhl, Christian. 'Autonomous Weapon Systems & Military AI: Cause Area Report'. Founders Pledge, May 2022. https://founderspledge.com/stories/autonomous-weapon-systems-and-military-artificial-intelligence-ai.

[436] Horowitz, Michael C, and Paul Scharre. 'AI and International Stability: Risks and Confidence-Building Measures'. Center for a New American Security, 12 January 2021. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

[437] Nye, Joseph S. 'Nuclear Learning and U.S.-Soviet Security Regimes'. *International Organization* 41, no. 3 (1987): 371–402. https://www.jstor.org/stable/2706750

[438] Gavin, Francis J. 'Strategies of Inhibition: U.S. Grand Strategy, the Nuclear Revolution, and Nonproliferation'. *International Security* 40, no. 1 (1 July 2015): 9–46. https://doi.org/10.1162/ISEC_a_00205.

[439] See generally: Koch, Lisa Langdon. 'Frustration and Delay: The Secondary Effects of Supply-Side Proliferation Controls'. *Security Studies* 28, no. 4 (8 August 2019): 773–806. https://doi.org/10.1080/09636412.2019.1631383.; Koch, Lisa Langdon. 'The NPT at 50 and the NSG at 43: How the Global Control of Nuclear Exports Has Slowed Proliferation'. *International History and Politics Newsletter*, Symposium on the 50th Anniversary of the Nuclear Non-Proliferation Treaty, 4, no. 1 (2018): 8–10. https://connect.apsanet.org/s34/wp-content/uploads/sites/19/2018/08/IHAP-Newsletter-4.1-Summer-2018.pdf

[440] Koch, Lisa Langdon. 'Military Regimes and Resistance to Nuclear Weapons Development'. *Security Studies* 0, no. 0 (10 May 2023): 1–32. https://doi.org/10.1080/09636412.2023.2197621. See generally Koch, Lisa. *Nuclear Decisions: Changing the Course of Nuclear Weapons Programs*. Oxford, New York: Oxford University Press, 2023.

[441] Maas, Matthijs M. 'How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons'. *Contemporary Security Policy* 40, no. 3 (6 February 2019): 285–311. https://doi.org/10.1080/13523260.2019.1576464.; Meer, Sico van der. 'Not That Bad: Looking Back on 65 Years of Nuclear Non-Proliferation Efforts'. Security and Human Rights 22, no. 1 (2011): 37–47. https://doi.org/10.1163/187502311796365862. Kaplow, Jeffrey M. Signing Away the Bomb: The Surprising Success of the Nuclear Nonproliferation Regime. Cambridge: Cambridge University Press, 2022. https://doi.org/10.1017/9781009216746.; Robichaud, Carl. 'The Puzzle of Non-Proliferation'. Asterisk, June 2023. https://asteriskmag.com/issues/03/the-puzzle-of-non-proliferation.

[442] Meer, Sico van der. 'Forgoing the Nuclear Option: States That Could Build Nuclear Weapons but Chose Not to Do So'. *Medicine, Conflict and Survival* 30, no. S1 (2014): s27–34. https://pubmed.ncbi.nlm.nih.gov/25175327/. Though for a different scoring, see also: Bleek, Philipp C. 'When Did (and Didn't) States Proliferate? Chronicling the Spread of Nuclear Weapons'. Belfer Center for Science and International Affairs, 2017. https://www.belfercenter.org/sites/default/files/files/publication/When%20Did%20%28and%20Didn%27t%29%20States%20Proliferate%3F_1.pdf.

→ General design lessons from existing treaty regimes: drawing insights from the design and efficacy of a range of treaties—including the Single Convention on Narcotic Drugs (SCND), the Vienna Convention on Psychotropic Substances (VCPS), the Convention Against Illicit Trafficking of Narcotic Drugs and Psychotropic Substances (CAIT), the Montreal Protocol on Substances that Deplete the Ozone Layer, the Cartagena Protocol on Biosafety to the Convention on Biological Diversity, the Biological Weapons Convention (BWC), the Treaty on the Non-Proliferation of Nuclear Weapons (NPT), the Convention on Nuclear Safety, the Convention on International Trade in Endangered Species (CITES), the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal, and the Bern Convention on the Conservation of European Wildlife and Natural Habitats—to derive design lessons for a global regulatory system dedicated to the regulation of safety concerns from high-risk AI;[443]

→ Effective use of international access and benefit distribution mechanisms in conjunction with proliferation control measures: the efficacy of the IAEA's "dual mandate" to enable the transfer of peaceful nuclear technology whilst seeking to curtail its use for military purposes;[444]

→ Effective monitoring and verification (M&V) mechanisms in arms control regimes: M&V implementation across three types of nuclear arms control treaties: nonproliferation treaties, US-USSR/Russia arms limitation treaties, and nuclear test bans;[445]

→ Scientific community (temporary) moratoria on research: the Asilomar Conference[446] and the H5N1 gain-of-function debate;[447]

→ Instances where treaty commitments, institutional infighting, or bureaucratic politics contributed to technological restraint: a range of cases resulting in cancellation of weapon systems development, including nuclear-ramjet powered cruise missiles, "continent killer" nuclear warheads, nuclear-powered aircraft, "death dust" radiological weapons, various types of anti-ballistic-missile defense, and many others.[448]

→ International institutional design lessons from successes and failures in other areas: global governance successes and failures in the regime complexes for environment, security, and/or trade;[449]

→ Successful use of soft-law governance tools for shaping emerging technologies: Internet Corporation for Assigned Names and Numbers, Motion Picture Association of America, Federal Trade

[443] See the framework set out in: Llerena, Stephan. 'Global Governance of High-Risk Artificial Intelligence', 27 October 2023. (draft manuscript).

[444] Law, Harry, and Lewis Ho. 'Can a Dual Mandate Be a Model for the Global Governance of AI?' *Nature Reviews Physics*, 27 October 2023, 1–2. https://doi.org/10.1038/s42254-023-00670-4.; see also Stafford, Eoghan, and Robert F Trager. 'The IAEA Solution: Knowledge Sharing to Prevent Dangerous Technology Races'. Centre for the Governance of AI, 2022. https://www.governance.ai/research-paper/knowledge-sharing-to-prevent-dangerous-technology-races.

[445] Baker, Mauricio. 'Nuclear Arms Control Verification and Lessons for AI Treaties'. arXiv, 8 April 2023. http://arxiv.org/abs/2304.04123.

[446] Grace, Katja. 'The Asilomar Conference: A Case Study in Risk Mitigation'. Technical Report. Berkeley, CA: Machine Intelligence Research Institute, 15 July 2015. https://intelligence.org/files/TheAsilomarConference.pdf.

[447] Wang, Jasmine. 'What the AI Community Can Learn From Sneezing Ferrets and a Mutant Virus Debate'. *Partnership on AI* (blog), 9 December 2020. https://medium.com/partnership-on-ai/lessons-for-the-ai-community-from-the-h5n1-controversy-32432438a82e

[448] Maas, Matthijs. 'Paths Untaken: The History, Epistemology and Strategy of Technological Restraint, and Lessons for AI'. *Verfassungsblog* (blog), 9 August 2022. https://verfassungsblog.de/paths-untaken/.

[449] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Should Artificial Intelligence Governance Be Centralised?: Design Lessons from History'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–34. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375857. Roberts, Huw, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. 'Global AI Governance: Barriers and Pathways Forward'. SSRN Scholarly Paper. Rochester, NY, 29 September 2023. https://doi.org/10.2139/ssrn.4588040.

Commission consent decrees, Federal Communications Commission's power over broadcaster licensing, Entertainment Software Rating Board, NIST Framework for Improving Critical Infrastructure Cybersecurity, Asilomar rDNA Guidelines, International Gene Synthesis Consortium, International Society for Stem Cell Research Guidelines, BASF Code of Conduct, Environmental Defense Fund, and DuPont Risk Framework;[450]

→ Successful use of participatory mechanisms in improving risk assessment: use of scenario methods and risk assessments in climate impact research.[451]

4.4. Lessons derived from ethics and political theory

Mapping the space of principles or criteria for "ideal AI governance":[452]

→ Mapping broad normative desiderata for good governance regimes for advanced AI,[453] either in terms of outputs or in terms of process;[454]

→ Understanding how to weigh different good outcomes post-TAI-deployment;[455]

→ Understanding the different functional goals and tradeoffs in good international institutional design.[456]

II. Option-identifying work: Mapping actors and affordances

Strategic clarity requires an understanding not just of the features of the advanced AI governance problem, but also of the options in response.

This entails mapping the range of possible levers that could be used in response to this problem. Critically, this is not just about speculating about what governance tools we may want to put in place for future advanced AI systems mid-transition (after they have arrived). Rather, there might be actions we could take in the "pre-emergence" stage to adequately prepare ourselves.[457]

---

[450] Gutierrez, Carlos Ignacio, Gary E. Marchant, and Lucille Tournas. 'Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance'. *JURIMETRICS* 61, no. 1 (29 December 2020). https://doi.org/10.2139/ssrn.3775271.

[451] Hollis, Helena, and Jess Whittlestone. 'Participatory AI Futures: Lessons from Research in Climate Change'. *Medium* (blog), 24 August 2021. https://medium.com/@helena.hollis.14/participatory-ai-futures-lessons-from-research-in-climate-change-34e3580553f8.

[452] The term is originally from: Dafoe, Allan. 'AI Governance: A Research Agenda'. Oxford: Center for the Governance of AI, Future of Humanity Institute, 2018. https://www.fhi.ox.ac.uk/govaiagenda/.

[453] Bostrom, Nick, Allan Dafoe, and Carrick Flynn. 'Public Policy and Superintelligent AI: A Vector Field Approach'. In *Ethics of Artificial Intelligence*, edited by S.M. Liao. Oxford University Press, 2019. http://www.nickbostrom.com/papers/aipolicy.pdf. See also: RoryG. 'What If AI Development Goes Well?'. EA Forum, 3 August 2022. https://forum.effectivealtruism.org/posts/9EjMoD8BRhXEsfzMh/what-if-ai-development-goes-well-3.

[454] Erman, Eva, and Markus Furendal. 'Artificial Intelligence and the Political Legitimacy of Global Governance'. *Political Studies*, 3 October 2022, 00323217221126665. https://doi.org/10.1177/00323217221126665.

[455] Karnofsky, Holden. 'Important, Actionable Research Questions for the Most Important Century'. EA Forum, 24 February 2022. https://forum.effectivealtruism.org/posts/zGiD94SHwQ9MwPyfW/important-actionable-research-questions-for-the-most.

[456] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890.

[457] The terms "mid-transition" and "pre-emergence" are drawn from: Manheim, David. 'AI Governance across Slow/Fast Takeoff and Easy/Hard Alignment Spectra'. AI Alignment Forum, 3 April 2022. https://www.alignmentforum.org/posts/xxMYFKLqiBJZRNoPj/ai-governance-across-slow-fast-takeoff-and-easy-hard.

Within the field, there has been extensive work on options and areas of intervention. Yet there is no clear, integrated map of the advanced AI governance landscape and its gaps. Sam Clarke proposes that there are different ways of carving up the landscape, such as based on different types of interventions, different geographic hubs, or "Theories of Victory."[458] To extend this, one might segment the advanced AI governance solution space along work which aims to identify and understand, in turn:[459]

→ Key actors that will likely (be in a strong position to) shape advanced AI;

→ Levers of influence (by which these actors might shape advanced AI);

→ Pathways towards influencing these actors to deploy their levers well.[460]

1. Potential key actors shaping advanced AI

In other words, whose decisions might especially affect the development and deployment of advanced AI, directly or indirectly, such that these decisions should be shaped to be as beneficial as possible?

> Key actors can be defined as "actors whose *key decisions* will have significant impact on shaping the *outcomes from advanced AI*, either directly (first-order), or by strongly affecting such decisions made by other actors (second-order)."[461]
>
> Key decisions can be further defined as "a choice or series of choices by a key actor to use its *levers of governance*, in ways that directly affect *beneficial advanced AI outcomes*, and which are hard to reverse."[462]

Some work in this space explores the relative importance of (the decisions of) different types of key actors:

→ The roles of state vs. firms vs. AI researchers in shaping AI policy;[463]

[458] Clarke, Sam. 'The Longtermist AI Governance Landscape: A Basic Overview'. EA Forum, 18 January 2022. https://forum.effectivealtruism.org/posts/ydpo7LcJWhrr2GJrx/the-longtermist-ai-governance-landscape-a-basic-overview.
[459] For definitions of these terms, see the start of each subsection and also: Maas, Matthijs, 'Concepts in advanced AI governance: A literature review of key terms and definitions.' *Institute for Law & AI*. AI Foundations Report 3. (October 2023). https://www.legalpriorities.org/research/advanced-ai-gov-concepts
[460] Two notes are relevant here: for one, in this section I bucket research horizontally along the categories of "key actors," "levers" (of these actors), and "pathways to influence" (on these actors). However, in practice, many specific analyses of interventions would integrate these three—discussing each actor both in terms of the levers available to it as well as pathways by which their decisions might be informed. In the second place, in this model different pathways are mostly treated as being actor-specific (that is, bucketed by which actor they are meant to influence). In some cases, however, we might consider that some pathways might be lever-specific (e.g., some types of advocacy are more suited to prompting the use of some types of government action than others are). I thank Suzanne van Arsdale for pointing out this distinction.
[461] For discussion of these terms, see also Maas, Matthijs, 'Concepts in advanced AI governance: A literature review of key terms and definitions.' *Institute for Law & AI*. AI Foundations Report 3. (October 2023). https://www.legalpriorities.org/research/advanced-ai-gov-concepts (discussing various technical, policy, and strategy-focused definitions of this field).
[462] Ibid.
[463] Leung, Jade. 'Who Will Govern Artificial Intelligence? Learning from the History of Strategic Politics in Emerging Technologies'. University of Oxford, 2019. https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665.

- → Role of "epistemic communities" of scientific experts,[464] especially members of the AI research community;[465]
- → The role of different potentially relevant stakeholders for responsible AI systems across its development chain, from individual stakeholders to organizational stakeholders to national/international stakeholders;[466]
- → The relative role of expert advice vs. public pressure in shaping policymakers' approach to AI;[467]
- → Role of different actors in and around the corporation in shaping lab policy,[468] including actors within the lab (e.g., senior management, shareholders, AI lab employees, and employee activists)[469] and actors outside the lab (e.g., corporate partners and competitors, industry consortia, nonprofit organizations, the public, the media, and governments).[470]

Other work focuses more specifically on mapping particular key actors whose decisions may be particularly important in shaping advanced AI outcomes, depending on one's view of strategic parameters.

The following list should be taken more as a "landscape" review than a literature review, since coverage of different actors differs amongst papers. Moreover, while the list aims to be relatively inclusive of actors, it is clear that the (absolute and relative) importance of each of these actors obviously differs hugely between worldviews and approaches.

1.1. AI developer (lab & tech company) actors

Leading AI firms pursuing AGI:

- → OpenAI,

[464] Maas, Matthijs M. 'How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons'. *Contemporary Security Policy* 40, no. 3 (6 February 2019): 285–311. https://doi.org/10.1080/13523260.2019.1576464.

[465] Shevlane, Toby. 'The Artefacts of Intelligence: Governing Scientists' Contribution to AI Proliferation'. University of Oxford, 2022. https://www.governance.ai/research-paper/the-artefacts-of-intelligence-governing-scientists-contribution-to-ai-proliferation.

[466] Deshpande, Advait, and Helen Sharp. 'Responsible AI Systems: Who Are the Stakeholders?' In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 227–36. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3514094.3534187.

[467] Schiff, Daniel S. 'Setting the Agenda for AI: Actors, Issues, and Influence in United States Artificial Intelligence Policy'. Georgia Institute of Technology, 2022. https://osf.io/kw8xd/. (in the US context).

[468] Cihon, Peter, Jonas Schuett, and Seth D. Baum. 'Corporate Governance of Artificial Intelligence in the Public Interest'. *Information* 12, no. 7 (July 2021): 275. https://doi.org/10.3390/info12070275. (discussing actors both "inside the corporation—managers, workers, and investors—and outside the corporation—corporate partners and competitors, industry consortia, nonprofit organizations, the public, the media, and governments"); Baum, Seth, and Jonas Schuett. 'The Case for Long-Term Corporate Governance of AI'. Effective Altruism Forum, 3 November 2021. https://forum.effectivealtruism.org/posts/5MZpxbJJ5pkEBpAAR/the-case-for-long-term-corporate-governance-of-ai.; Leung, Jade. 'Why Companies Should Be Leading on AI Governance'. 16 May 2019. https://forum.effectivealtruism.org/posts/fniRhiPYw8b6FETsn/jade-leung-why-companies-should-be-leading-on-ai-governance.

[469] Belfield, Haydn. 'Activism by the AI Community: Analysing Recent Achievements and Future Prospects'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 15–21. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375814. And see generally: Nedzhvetskaya, Nataliya, and J. S. Tan. 'The Role of Workers in AI Ethics and Governance'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press. Accessed 21 October 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.68.

[470] Cihon, Peter, Jonas Schuett, and Seth D. Baum. 'Corporate Governance of Artificial Intelligence in the Public Interest'. *Information* 12, no. 7 (July 2021): 275. https://doi.org/10.3390/info12070275.

- → DeepMind,
- → Anthropic,
- → Aleph Alpha,
- → Adept,
- → Cohere,
- → Inflection,
- → Keen,
- → xAI.[471]

Chinese labs and institutions researching "general AI";

- → Baidu Research,
- → Alibaba DAMO Academy,
- → Tencent AI Lab,
- → Huawei,
- → JD Research Institute,
- → Beijing Institute for General Artificial Intelligence;
- → Beijing Academy of Artificial Intelligence, etc.[472]

Large tech companies[473] that may take an increasingly significant role in AGI research:

- → Microsoft,
- → Google,
- → Facebook,
- → Amazon.

Future frontier labs, currently not known but to be established/achieve prominence (e.g., "Magma"[474]).

1.2. AI services & compute hardware supply chains

AI services supply chain actors:[475]

[471] Listed in: Hogarth, Ian. 'We Must Slow down the Race to God-like AI'. *Financial Times*, 13 April 2023. https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2. See also informally: Spencer, Michael, and Charlie Guo. 'The Top Six Rivals Competing with OpenAI'. Substack newsletter. *AI Supremacy* (blog), 27 April 2023. https://aisupremacy.substack.com/p/the-top-six-rivals-competing-with?publication_id=396235. For older overviews of the landscape of labs pursuing AGI, see: Fitzgerald, McKenna, Aaron Boddy, and Seth D. Baum. '2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy'. Global Catastrophic Risk Institute Technical Report. Global Catastrophic Risk Institute, 2020. https://gcrinstitute.org/papers/055_agi-2020.pdf. And previously: Baum, Seth. 'A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy'. Global Catastrophic Risk Institute Technical Report. Global Catastrophic Risk Institute, 12 November 2017. https://papers.ssrn.com/abstract=3070741.

[472] Hannas, William, Huey-Meei Chang, Daniel Chou, and Brian Fleeger. 'China's Advanced AI Research: Monitoring China's Paths to "General" Artificial Intelligence'. Center for Security and Emerging Technology, July 2022. https://cset.georgetown.edu/publication/chinas-advanced-ai-research/.

[473] See also: Leung, Jade. 'Why Companies Should Be Leading on AI Governance'. 16 May 2019. https://forum.effectivealtruism.org/posts/fniRhiPYw8b6FETsn/jade-leung-why-companies-should-be-leading-on-ai-governance.

[474] The particular name comes from: Cotra, Ajeya. 'Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover'. AI Alignment Forum, 18 July 2022. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.

[475] Cobbe, Jennifer, Michael Veale, and Jatinder Singh. 'Understanding Accountability in Algorithmic Supply Chains'. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1186–97, 2023. https://doi.org/10.1145/3593013.3594073.

- → Cloud computing providers:[476]
  - → Globally: Amazon Web Services (32%), Microsoft Azure (21%), and Google Cloud (8%); IBM;
  - → Chinese market: Alibaba, Huawei, and Tencent.

Hardware supply chain industry actors:[477]

- → Providers of optical components to photolithography machine manufacturers:
  - → Carl Zeiss AG [Germany], a key ASML supplier of optical lenses;[478]
- → Producers of extreme ultraviolet (EUV) photolithography machines:
  - → ASML [The Netherlands].[479]
- → Photoresist processing providers:
  - → Asahi Kasei and Tokyo Ohka Kogyo Co. [Japan].[480]
- → Advanced chip manufacturing:
  - → TMSC [Taiwan];
  - → Intel [US];
  - → Samsung [South Korea].
- → Semiconductor intellectual property owners and chip designers:
  - → Arm [UK];
  - → Graphcore [UK].
- → DRAM integrated circuit chips:
  - → Samsung (market share 44%) [South Korea];
  - → SK hynix (27%) [South Korea];
  - → Micron (22%) [US].

---

[476] Belfield, Haydn, and Shin-Shin Hua. 'Compute and Antitrust: Regulatory implications of the AI hardware supply chain, from chip design to cloud APIs'. *Verfassungsblog* (blog), 19 August 2022. https://verfassungsblog.de/compute-and-antitrust/.

[477] For an interactive supply chain exploration tool, see: Emerging Technology Observatory. 'Supply Chain Explorer: Advanced Chips', 16 October 2022. https://chipexplorer.eto.tech/.; for overviews, see: See Khan, Saif. 'The Semiconductor Supply Chain: Assessing National Competitiveness'. Center for Security and Emerging Technology, January 2021. https://cset.georgetown.edu/publication/the-semiconductor-supply-chain/.; Belfield, Haydn, and Shin-Shin Hua. 'Compute and Antitrust: Regulatory implications of the AI hardware supply chain, from chip design to cloud APIs'. *Verfassungsblog* (blog), 19 August 2022. https://verfassungsblog.de/compute-and-antitrust/.; Elmgren, Karson. 'China's Silicon Future'. Asterisk, November 2022. https://asteriskmag.com/issues/1/china-s-silicon-future.

[478] Raaijmakers, René. *ASML's Architects: The Story of the Engineers Who Shaped the World's Most Powerful Chip Machines*. Nijmegen: Techwatch Books, 2019.; see also: Tung, An-Chi, and Henry Wan. 'Organisational Investment: The Case of ASML—Can the Product Make the Producer?' *Foreign Trade Review* 58, no. 1 (1 February 2023): 176–91. https://doi.org/10.1177/00157325221127606. ASML. 'ZEISS and ASML Strengthen Partnership for Next Generation of EUV Lithography', 2016. https://www.asml.com/en/news/press-releases/2016/zeiss-and-asml-strengthen-partnership-for-next-generation-of-euv-lithography. Hoyng Rokh Monegier. 'ASML and Carl Zeiss SMT v. Nikon – Immersion Lithography', 28 July 2020. https://www.hoyngrokhmonegier.com/news-insights/case-studies/asml-and-carl-zeiss-smt-v-nikon-immersion-lithography/.

[479] Belfield, Haydn, and Shin-Shin Hua. 'Compute and Antitrust: Regulatory implications of the AI hardware supply chain, from chip design to cloud APIs'. *Verfassungsblog* (blog), 19 August 2022. https://verfassungsblog.de/compute-and-antitrust/.; See also: Yglesias, Matthew. 'At Last, an AI Existential Risk Policy Idea'. Slow Boring, 28 September 2022. https://www.slowboring.com/p/at-last-an-ai-existential-risk-policy.

[480] Elmgren, Karson. 'China's Silicon Future'. Asterisk, November 2022. https://asteriskmag.com/issues/1/china-s-silicon-future.

- → GPU providers:
  - → Intel (market share 62%) [US];
  - → AMD (18%) [US];
  - → NVIDIA (20%) [US].

1.3. AI industry and academic actors

Industry bodies:

- → Partnership on AI;
- → Frontier Model Forum;[481]
- → ML Commons;[482]
- → IEEE (Institute of Electrical and Electronics Engineers) + IEEE-SA (standards body);
- → ISO (and IEC).

Standard-setting organizations:

- → US standard-setting organizations (NIST);
- → European Standards Organizations (ESOs), tasked with setting standards for the EU AI Act: the European Committee for Standardisation (CEN), European Committee for Electrotechnical Standardisation (CENELEC), and European Telecommunications Standards Institute (ETSI);[483]
- → VDE (influential German standardization organization).[484]

Software tools & community service providers:

- → arXiv;
- → GitHub;
- → Colab;
- → Hugging Face.

Academic communities:

- → Scientific ML community;[485]
- → AI conferences: NeurIPS, AAAI/ACM, ICLR, IJCAI-ECAI, AIES, and FAccT, etc.;
- → AI ethics community and various subcommunities;
- → Numerous national-level academic or research institutes.

Other active tech community actors:

---

[481] Google. 'A New Partnership to Promote Responsible AI'. Google, 26 July 2023. https://blog.google/outreach-initiatives/public-policy/google-microsoft-openai-anthropic-frontier-model-forum/.

[482] MLCommons. 'MLCommons - Philosophy'. Accessed 2 December 2022. https://mlcommons.org/.

[483] O'Keefe, Cullen, Jade Leung, and Markus Anderljung. 'How Technical Safety Standards Could Promote TAI Safety'. Effective Altruism Forum, 8 August 2022. https://forum.effectivealtruism.org/posts/zvbGXCxc5jBowCuNX/how-technical-safety-standards-could-promote-tai-safety.

[484] VDE. 'VDE and Partners Develop Quality Standards for AI Test and Training Data'. VDE, 10 February 2022. https://www.vde.com/ai-training-data.

[485] Shevlane, Toby. 'The Artefacts of Intelligence: Governing Scientists' Contribution to AI Proliferation'. University of Oxford, 2022. https://www.governance.ai/research-paper/the-artefacts-of-intelligence-governing-scientists-contribution-to-ai-proliferation.; Prunkl, Carina E. A., Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 'Institutionalizing Ethics in AI through Broader Impact Requirements'. *Nature Machine Intelligence* 3, no. 2 (February 2021): 104–10. https://doi.org/10.1038/s42256-021-00298-y.

- → Open-source machine learning software community;[486]
- → "Open"/diffusion-encouraging[487] AI community (e.g., Stability.ai, Eleuther.ai);[488]
- → Hacker communities;
- → Cybersecurity and information security expert communities.[489]

1.4. State and governmental actors

Various states, and their constituent (government) agencies or bodies that are, plausibly will be, or potentially could be moved to be in powerful positions to shape the development of advanced AI.

The United States

Key actors in the US:[490]

- → Executive Branch actors;[491]
- → Legislative Branch;[492]
- → Judiciary;[493]

[486] Langenkamp, Max, and Daniel N. Yue. 'How Open Source Machine Learning Software Shapes AI'. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 385–95. Oxford United Kingdom: ACM, 2022. https://doi.org/10.1145/3514094.3534167.; Engler, Alex. 'How Open-Source Software Shapes AI Policy'. *Brookings* (blog), 10 August 2021. https://www.brookings.edu/research/how-open-source-software-shapes-ai-policy/.

[487] I thank Di Cooke for suggesting this term. For an in-depth discussion and proposed alternate approaches, see: Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K Wei, Christoph Winter, et al. 'Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives'. Centre for the Governance of AI, 2023. https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.

[488] Phang, Jason, Herbie Bradley, Leo Gao, Louis Castricato, and Stella Biderman. 'EleutherAI: Going Beyond "Open Science" to "Science in the Open"'. arXiv, 12 October 2022. https://doi.org/10.48550/arXiv.2210.06413. As well as the overview at AndreFerretti, and mic. 'Navigating the Open-Source AI Landscape: Data, Funding, and Safety'. EA Forum, 12 April 2023. https://forum.effectivealtruism.org/posts/N25EARxvbxYJa5pbB/navigating-the-open-source-ai-landscape-data-funding-and; though for critical views of actors in this community, see: Seger, Elizabeth. 'What Do We Mean When We Talk About "AI Democratisation"?' *GovAI Blog* (blog), 7 February 2023. https://www.governance.ai/post/what-do-we-mean-when-we-talk-about-ai-democratisation.

[489] Ladish, Jeffrey. 'Information Security Considerations for AI and the Long Term Future'. EA Forum, 2 May 2022. https://forum.effectivealtruism.org/posts/WqQDCCLWbYfFRwubf/information-security-considerations-for-ai-and-the-long-term.; Zabel, Claire, and Luke Muelhauser. 'Information Security Careers for GCR Reduction'. EA Forum, 21 June 2019. https://forum.effectivealtruism.org/posts/ZJiCfwTy5dC4CoxqA/information-security-careers-for-gcr-reduction.

[490] Bowerman, Niel. 'The Case for Building Expertise to Work on US AI Policy'. 80,000 Hours, December 2020. https://80000hours.org/articles/us-ai-policy/. I especially thank Di Cooke and Carlos Ignacio Gutierrez for some suggestions with regards to US actors.

[491] These may include, but are not limited to: President; Office of Science and Technology Policy (OSTP), especially its National AI Initiative Office (NAIIO); National Security Council (NSC); and National Science Foundation (basic and applied grants). National Science Foundation. 'Artificial Intelligence (AI) at NSF'. Accessed 20 February 2023. https://www.nsf.gov/cise/ai.jsp.

[492] These may include, but are not limited to: various Congressional actors, such as intelligence committees; appropriations committees; commerce committees; Congressional AI Caucus; and House Committee on Science, Space, and Technology.

[493] Scherer, Matthew U. 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies'. *Harvard Journal of Law & Technology*, no. 2 (Spring 2016). http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf. Deeks, Ashley. 'The Judicial Demand for Explainable Artificial Intelligence'. *Columbia Law Review* 119 (2019): 1829–50. https://pdfs.semanticscholar.org/bf49/b0a7dcf4c1af68cf80cc3fe4df60b95b0da4.pdf

- → Federal agencies;[494]
- → Intelligence community;[495]
- → Independent federal agencies;[496]
- → Relevant state and local governments, such as the State of California (potentially significant extraterritorial regulatory effects),[497] State of Illinois and State of Texas (among first states to place restrictions on biometrics), etc.

China

Key actors in China:[498]

- → 20th Central Committee of the Chinese Communist Party;

[494] These may include, but are not limited to: Department of Justice (DoJ); Department of Commerce, including the National Institute of Standards and Technology (NIST); Office of Management and Budget (OMB); Bureau of Industry and Security; Department of Defense (DoD), including the CDAO (Chief Digital and Artificial Intelligence Office); DARPA (Defense Advanced Research Projects Agency); Emerging Capabilities Policy Office; Office of Net Assessment; National Security Commission on Artificial Intelligence (completed); The Department of Homeland Security, including FEMA or the US Customs and Border Protection (use of facial recognition); Department of Health and Human Services, including the Food and Drug Administration (FDA) (for approving medical AI systems); Department of Labor; and Department of Energy.
For work on some of these, see amongst others: Barrett, Anthony M., Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. 'Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks'. arXiv, 17 June 2022. https://doi.org/10.48550/arXiv.2206.08966.; National Security Commission on Artificial Intelligence. 'Final Report'. National Security Commission on Artificial Intelligence, March 2021. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf. ; 'U.S. Department of Homeland Security Artificial Intelligence Strategy'. U.S. Department of Homeland Security, 3 December 2020. https://www.dhs.gov/publication/us-department-homeland-security-artificial-intelligence-strategy. See generally also Weaver, John Frank. 'Regulation of Artificial Intelligence in the United States'. In *Research Handbook on the Law of Artificial Intelligence*, 155–212. Edward Elgar Publishing, 2018. https://www.elgaronline.com/display/edcoll/9781786439048/9781786439048.00018.xml.

[495] These may include, but are not limited to: IARPA, ODNI, and In-Q-Tel.

[496] These may include, but are not limited to: the Federal Trade Commission (FTC), able to regulate a broad range of harms of AI systems and to seek injunctions to order a company to cease certain unfair or deceptive practices; the Securities and Exchange Commission, for shaping financial applications of AI; the Committee on Foreign Investment in the United States (CFIUS), a federal interagency committee able to review foreign investments in US companies on national security grounds.
On the FTC's role in AI governance, see: Selbst, Andrew D., and Solon Barocas. 'Unfair Artificial Intelligence: How FTC Intervention Can Overcome the Limitations of Discrimination Law'. SSRN Scholarly Paper. Rochester, NY, 8 August 2022. https://papers.ssrn.com/abstract=4185227.; and Okerlund, Johanna, Evan Klasky, Aditya Middha, Sujin Kim, Hannah Rosenfeld, Molly Kleinman, and Shobita Parthasarathy. 'What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them'. Ford School of Public Policy, University of Michigan, 2022. https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox. Spiro, Michael. 'The FTC and AI Governance: A Regulatory Proposal'. *Seattle Journal of Technology, Environmental & Innovation Law* 10, no. 1 (19 December 2020). https://digitalcommons.law.seattleu.edu/sjteil/vol10/iss1/2.; Casper, Stephen, Phillip Christoffersen, and Rui-Jie Yew. 'The Slippery Slope from DALLE-2 to Deepfake Anarchy'. Effective Altruism Forum, 5 November 2022. https://forum.effectivealtruism.org/posts/Bnp9YDqErNXHmTvvE/the-slippery-slope-from-dalle-2-to-deepfake-anarchy.; FTC. 'FTC Launches New Office of Technology to Bolster Agency's Work'. Federal Trade Commission, 16 February 2023. https://www.ftc.gov/news-events/news/press-releases/2023/02/ftc-launches-new-office-technology-bolster-agencys-work.

[497] See e.g. Josephson, Henry. 'A California Effect for Artificial Intelligence', 2022. https://www.henryjos.com/p/a-california-effect-for-artificial.html.

[498] See generally Cheng, Jing, and Jinghan Zeng. 'Shaping AI's Future? China in Global AI Governance'. *Journal of Contemporary China* 0, no. 0 (8 August 2022): 1–17. https://doi.org/10.1080/10670564.2022.2107391.; Ding, Jeffrey. 'Deciphering China's AI Dream: The Context, Components, Capabilities, and Consequences of China's Strategy to Lead the World in AI'. Future of Humanity Institute, Governance of AI Program, March 2018. https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf?platform=hootsuite.; Roberts, Huw, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. 'The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation'. *AI & SOCIETY* 36, no. 1 (1 March 2021): 59–77. https://doi.org/10.1007/s00146-020-00992-2.

- → China's State Council;
- → Bureaucratic actors engaged in AI policy-setting;[499]
- → Actors and institutions engaged in track-II diplomacy on AI.[500]

The EU

Key actors in the EU:[501]

- → European Commission;
- → European Parliament;
- → Scientific research initiatives and directorates;[502]
- → (Proposed) European Artificial Intelligence Board and notified bodies.[503]

The UK

Key actors in the UK:[504]

- → The Cabinet Office;[505]

[499] These include: the Cyberspace Administration of China (CAC), Ministry of Industry and Information Technology, the Ministry of Science and Technology, and the National Science and Technology Ethics Committee; Standardization Administration of China (SAC). See also: Sheehan, Matt. 'China's New AI Governance Initiatives Shouldn't Be Ignored'. Carnegie Endowment for International Peace, 4 January 2022. https://carnegieendowment.org/2022/01/04/china-s-new-ai-governance-initiatives-shouldn-t-be-ignored-pub-86127. Zhou, Frank, and Summer Sun. 'How China Regulates Ethical Issues in "AI+life Science"'. International Bar Association, 21 October 2022. https://www.ibanet.org/china-regulates-ai-life-science.

[500] These include: the Institute for AI International Governance (I-AIIG) and Center for International Security and Strategy (CISS) at Tsinghua University, overseen by Madam Fu Ying (current Chairperson of the National People's Congress Foreign Affairs Committee). Ying, Fu, and John Allen. 'Together, The U.S. And China Can Reduce The Risks From AI'. *Noema*, 17 December 2020. https://www.noemamag.com/together-the-u-s-and-china-can-reduce-the-risks-from-ai.; U.S.-China Perception Monitor. 'Who Is Fu Ying?' Accessed 22 February 2023. https://uscnpm.org/who-is-fu-ying/.

[501] See generally: Stix, Charlotte. 'The European Artificial Intelligence Landscape'. Workshop Report. European Commission, 2018. https://ec.europa.eu/digital-single-market/en/news/european-artificial-intelligence-landscape.; Siegmann, Charlotte, and Markus Anderljung. 'The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market'. Centre for the Governance of AI, August 2022. https://www.governance.ai/research-paper/brussels-effect-ai. I thank Haydn Belfield, Jacob Arbeid, and Moritz Kleinalterkamp for suggestions and input.

[502] Including, but not limited to: the Joint Research Centre (JRC) and International Outreach for Human-Centric Artificial Intelligence Initiative (a joint initiative by the European Commission's Service for Foreign Policy Instruments (FPI) and the Directorate General for Communications Networks, Content and Technology (DG CONNECT), in collaboration with the European External Action Services (EEAS)). See European Commission. 'International Outreach for Human-Centric Artificial Intelligence Initiative', 2022. https://digital-strategy.ec.europa.eu/en/policies/international-outreach-ai.

[503] Stix, Charlotte. 'The Ghost of AI Governance Past, Present, and Future: AI Governance in the European Union'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press. Accessed 21 October 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.56.; and generally Stahl, Bernd Carsten, Rowena Rodrigues, Nicole Santiago, and Kevin Macnish. 'A European Agency for Artificial Intelligence: Protecting Fundamental Rights and Ethical Values'. *Computer Law & Security Review* 45 (1 July 2022): 105661. https://doi.org/10.1016/j.clsr.2022.105661.

[504] See Hadshar, Rose. 'Current UK Government Levers on AI Development'. EA Forum, 10 April 2023. https://forum.effectivealtruism.org/posts/BFBf5yPLoJMGozygE/current-uk-government-levers-on-ai-development. And also: Roberts, Huw, Alexander Babuta, Jessica Morley, Christopher Thomas, Mariarosaria Taddeo, and Luciano Floridi. 'Artificial Intelligence Regulation in the United Kingdom: A Path to Global Leadership?' SSRN Scholarly Paper. Rochester, NY, 1 September 2022. https://doi.org/10.2139/ssrn.4209504. For input on this section, I also thank Jess Whittlestone, Haydn Belfield, and Di Cooke.

[505] Including the Office of the Chief Scientific Advisor for National Security and Office for Science and Technology Strategy. For AI policies in the context of the UK National Resilience Strategy, see: Maas, Matthijs M., Diane Cooke, Tom Hobson, Lalitha Sundaram, Haydn Belfield, Lara Mani, Jess Whittlestone, and Seán Ó HÉigeartaigh. 'Reconfiguring Resilience for Existential Risk: Submission of Evidence to the Cabinet Office on the New UK National Resilience Strategy'. Centre for the Study of Existential Risk, 27 September 2021. https://www.cser.ac.uk/resources/reconfiguring-resilience-existential-risk-submission-evidence-cabinet-office-new-uk-national-resilience-strategy/.

- → Foreign Commonwealth and Development Office (FCDO);
- → Ministry of Defence (MoD);[506]
- → Department for Science, Innovation and Technology (DSIT);[507]
- → UK Parliament;[508]
- → The Digital Regulators Cooperation Forum;
- → Advanced Research and Invention Agency (ARIA).

Other states with varying roles

Other states that may play key roles because of their general geopolitical influence, AI-relevant resources (e.g., compute supply chain and significant research talent), or track record as digital norm setters:

- → Influential states: India, Russia, and Brazil;
- → Significant AI research talent: France, and Canada;
- → Hosting nodes in the global hardware supply chain: US (NVIDIA), Taiwan (TSMC), South Korea (Samsung), the Netherlands (ASML), Japan (photoresist processing), UK (Arm), and Germany (Carl Zeiss AG);
- → Potential (regional) neutral hubs: Singapore[509] and Switzerland;[510]
- → Global South coalitions: states from the Global South[511] and coalitions of Small Island Developing States (SIDS);[512]
- → Track record of (digital) norm-setters: Estonia and Norway.[513]

[506] Including actors such as: Defence AI and Autonomy Unit (DAU) (strategy level policy across UK Defence), Defence AI Centre (DAIC) (unit of excellence for AI best practices and guidance across UK Defence), and Defence Science and Technology Laboratory (DSTL).

[507] Within which sit: the Government Office for Science, Office for Science and Technology Strategy, Office for Artificial Intelligence, and the Frontier AI Taskforce.

[508] Including: (formerly) The House of Lords Select Committee on AI, Commons Science and Technology Committee, and AI APPG.

[509] Chua, Yi-Yang. 'Singapore AI Policy Career Guide'. EA Forum, 21 January 2021. https://forum.effectivealtruism.org/posts/umeMcbD4jDseLjsgT/singapore-ai-policy-career-guide.

[510] Fischer, Sophie-Charlotte, and Andreas Wenger. 'A Politically Neutral Hub for Basic AI Research'. Policy Perspectives. Zurich: CSS, ETH Zurich, March 2019. http://www.css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP7-2_2019-E.pdf.

[511] Adan, Sumaya Nur. 'The Case for Including the Global South in AI Governance Discussions'. GovAI Blog, 20 October 2023. https://www.governance.ai/post/the-case-for-including-the-global-south-in-ai-governance-conversations.; Png, Marie-Therese. 'At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance'. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1434–45. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3531146.3533200.; Garcia, Eugenio V. 'The Militarization of Artificial Intelligence: A Wake-up Call for the Global South', September 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3452323 . See also: Abungu, Cecil, Michelle Malonza, and Sumaya Nur Adan. 'Can Apparent Bystanders Distinctively Shape An Outcome? The Extent To Which Some Global South Countries Could Matter in the Global Catastrophic Risk-Focused Governance of Artificial Intelligence Development'. ILINA STAI Paper, 2023 (forthcoming).

[512] Estier, Malou, Belinda Cleeland, and Maxime Stauffer. 'Safe and Beneficial Artificial Intelligence for Small-Island Developing States'. Simon Institute for Longterm Governance, 25 July 2023. https://www.simoninstitute.ch/blog/post/safe-and-beneficial-artificial-intelligence-for-small-island-developing-states/.

[513] See generally: Andersen, Philip Hall, Henrik Øberg Myhre, Andreas Massey, Jakob Graabak, Sanna Baug Warholm, and Erik Aunvåg Matsen. 'Why Scale Is Overrated: The Case for Increasing EA Policy Efforts in Smaller Countries'. EA Forum, 15 August 2021. https://forum.effectivealtruism.org/posts/7SjtFYo6sCe3588Tx/why-scale-is-overrated-the-case-for-increasing-ea-policy.

1.5. Standard-setting organizations

International standard-setting institutions:[514]

- → ISO;
- → IEC;
- → IEEE;
- → CEN/CENELEC;
- → VDE (Association for Electrical, Electronic & Information Technologies) and its AI Quality & Testing Hub.[515]

1.6. International organizations

Various United Nations agencies:[516]

- → ITU;[517]
- → UNESCO;[518]
- → Office of the UN Tech Envoy (conducting the process leading to the Global Digital Compact in 2024);
- → UN Science, Technology, and Innovation (STI) Forum;
- → UN Executive Office of the Secretary-General;
- → UN General Assembly;
- → UN Security Council (UNSC);
- → UN Human Rights Council;[519]
- → Office of the High Commissioner on Human Rights;[520]
- → UN Chief Executives Board for Coordination;[521]
- → Secretary-General's High-Level Advisory Board on Effective Multilateralism (HLAB);

---

[514] Cihon, Peter. 'Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development'. Technical Report. Oxford: Center for the Governance of AI, Future of Humanity Institute, University of Oxford, April 2019. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.; Lorenz, Philippe. 'AI Governance through Political Fora and Standards Developing Organizations'. Stiftung Neue Verantwortung, September 2020. https://www.stiftung-nv.de/de/publikation/ai-governance-through-political-fora-and-standards-developing-organizations.

[515] 'Artificial Intelligence Put to Test: State of Hesse and VDE Present First AI Quality & Testing Hub Nationwide', 13 February 2023. https://www.vde.com/en/press/press-releases/2023-02-13-pk-aiq-hub.

[516] Garcia, Eugenio V. 'Multilateralism and Artificial Intelligence: What Role for the United Nations?' In *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, 18. Boca Raton: CRC Press, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3779866.; see also Sepasspour, Rumtin. 'A Reality Check and a Way Forward for the Global Governance of Artificial Intelligence'. *Bulletin of the Atomic Scientists*, 10 September 2023. https://www.tandfonline.com/doi/abs/10.1080/00963402.2023.2245249. And previously, Nindler, Reinmar. 'The United Nation's Capability to Manage Existential Risks with a Focus on Artificial Intelligence'. *International Community Law Review* 21, no. 1 (11 March 2019): 5–34. https://doi.org/10.1163/18719732-12341388.; and see the overview at: Kunz, Martina. 'AI and International Organizations'. Accessed 31 October 2022. https://globalaigov.org/participants/igos.html. I thank Eugenio Vargas Garcia for additional suggestions.

[517] ITU. 'United Nations Activities on Artificial Intelligence (AI) 2019'. ITU, 2019. https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf.

[518] Garcia, Eugenio V. 'Multilateralism and Artificial Intelligence: What Role for the United Nations?' In *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, 18. Boca Raton: CRC Press, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3779866.; see also: UNESCO. 'UNESCO Member States Adopt the First Ever Global Agreement on the Ethics of Artificial Intelligence'. UNESCO, 25 November 2021. https://en.unesco.org/news/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence.

[519] European Center for Not-for-Profit Law. 'UN HRC 51: New and Emerging Technologies and Human Rights at the Heart of New Resolutions Adopted', 13 October 2022. https://ecnl.org/news/un-hrc-51-new-and-emerging-technologies-and-human-rights-heart-new-resolutions-adopted.

[520] UN OHCHR. 'New and Emerging Digital Technologies and Human Rights'. OHCHR. Accessed 30 January 2023. https://www.ohchr.org/en/hr-bodies/hrc/advisory-committee/digital-technologiesand-hr.

[521] UN - CEB. 'Artificial Intelligence', 2020. https://unsceb.org/topics/artificial-intelligence.

→ Secretary-General's High-Level Advisory Body on Artificial Intelligence ("AI Advisory Body").[522]

Other international institutions already engaged on AI in some capacity[523] (in no particular order):

→ OECD;[524]
→ Global Partnership on AI;
→ G7;[525]
→ G20;[526]
→ Council of Europe (Ad Hoc Committee on Artificial Intelligence (CAHAI));[527]
→ NATO;[528]
→ AI Partnership for Defense;[529]
→ Global Road Traffic Forum;[530]
→ International Maritime Organisation;
→ EU-US Trade and Technology Council (TTC);[531]
→ EU-India Trade and Technology Council;
→ Multi-stakeholder fora: World Summit on the Information Society (WSIS), Internet Governance Forum (IGF), Global Summit on AI for Good,[532] and World Economic Forum (Centre for Trustworthy Technology).

Other international institutions not yet engaged on AI:

---

[522] Office of the Secretary-General's Envoy on Technology. 'High-Level Advisory Body on Artificial Intelligence'. United Nations, 2023. https://www.un.org/techenvoy/ai-advisory-body. United Nations. 'AI Advisory Body'. United Nations, 2023. https://www.un.org/en/ai-advisory-body.

[523] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890.; Schmitt, Lewin. 'Mapping Global AI Governance: A Nascent Regime in a Fragmented Landscape'. *AI and Ethics*, 17 August 2021. https://doi.org/10.1007/s43681-021-00083-y.

[524] OECD. 'State of Implementation of the OECD AI Principles: Insights from National AI Policies'. OECD Digital Economy Papers. OECD, 2021. https://www.oecd-ilibrary.org/content/paper/1cd40c44-en.

[525] See generally: Morin, Jean‐Frédéric, Hugo Dobson, Claire Peacock, Miriam Prys‐Hansen, Abdoulaye Anne, Louis Bélanger, Peter Dietsch, et al. 'How Informality Can Address Emerging Issues: Making the Most of the G7'. *Global Policy* 10, no. 2 (May 2019): 267–73. https://doi.org/10.1111/1758-5899.12668. (briefly discussing AI issues as an area where informal, like-minded club governance could excel).

[526] Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. 'Policy Brief: The Creation of a G20 Coordinating Committee for the Governance of Artificial Intelligence'. *AI and Ethics*, 6 October 2020. https://doi.org/10.1007/s43681-020-00019-y.

[527] Breuer, Marten. 'The Council of Europe as an AI Standard Setter'. *Verfassungsblog* (blog), 4 April 2022. https://verfassungsblog.de/the-council-of-europe-as-an-ai-standard-setter/.

[528] NATO. 'NATO's Data and Artificial Intelligence Review Board'. NATO, 13 October 2022. https://www.nato.int/cps/en/natohq/official_texts_208374.htm.; see also: Stanley-Lockman, Zoe, and Lena Trabucco. 'NATO's Role in Responsible AI Governance in Military Affairs'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press. Accessed 21 October 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.69.

[529] Trabucco, Lena. 'AI Partnership for Defense Is a Step in the Right Direction – But Will Face Challenges'. *Opinio Juris* (blog), 5 October 2020. http://opiniojuris.org/2020/10/05/ai-partnership-for-defense-is-a-step-in-the-right-direction-but-will-face-challenges/.

[530] Smith, Bryant Walker. 'New Technologies and Old Treaties'. *AJIL Unbound* 114 (ed 2020): 152–57. https://doi.org/10.1017/aju.2020.28.

[531] European Commission. 'EU-US Trade and Technology Council'. Accessed 30 January 2023. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/stronger-europe-world/eu-us-trade-and-technology-council_en. See also: O'Keefe, Cullen, Jade Leung, and Markus Anderljung. 'How Technical Safety Standards Could Promote TAI Safety'. Effective Altruism Forum, 8 August 2022. https://forum.effectivealtruism.org/posts/zvbGXCxc5jBowCuNX/how-technical-safety-standards-could-promote-tai-safety.

[532] Kunz, Martina. 'AI and Multi-Stakeholder Fora'. Accessed 31 October 2022. https://globalaigov.org/participants/fora.html.

→ International & regional courts: International Criminal Court (ICC), International Court of Justice (ICJ), and European Court of Justice.

1.7. Public, Civil Society, & media actors

Civil society organizations:[533]

→ Gatekeepers engaged in AI-specific norm-setting and advocacy: Human Rights Watch, Campaign to Stop Killer Robots,[534] and AlgorithmWatch;[535]
→ Civilian open-source intelligence (OSINT) actors engaged in monitoring state violations of human rights and international humanitarian law:[536] Bellingcat, NYT Visual Investigation Unit, CNS (Arms Control Wonk), Middlebury Institute, Forensic Architecture, BBC Africa Eye, Syrian Archive, etc.
→ Military AI mediation: Centre for Humanitarian Dialogue and Geneva Centre for Security Policy.[537]

Media actors:

→ Mass media;[538]
→ Tech media;
→ "Para-scientific media."[539]

Cultural actors:

→ Film industry (Hollywood, etc.);
→ Influential and widely read authors.[540]

2. Levers of governance (for each key actor)

That is, how might each key actor shape the development of advanced AI?

[533] On the relatively slow 2000s response to the threat of LAWS, see Carpenter, Charli. *'Lost' Causes, Agenda Vetting in Global Issue Networks and the Shaping of Human Security*. Ithaca: Cornell University Press, 2014. https://doi.org/10.7591/9780801470363.
[534] Campaign to Stop Killer Robots. 'About Us'. Accessed 5 September 2020. https://www.stopkillerrobots.org/about/. But see: Rosert, Elvira, and Frank Sauer. 'How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies'. *Contemporary Security Policy* 0, no. 0 (30 May 2020): 1–26. https://doi.org/10.1080/13523260.2020.1771508.
[535] See e.g. AlgorithmWatch. 'AI Ethics Guidelines Global Inventory'. *AlgorithmWatch* (blog), 2019. https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/.
[536] I thank Di Cooke for suggesting many of these.
[537] Geneva Centre for Security Policy. 'The Geneva Process on AI Principles'. Accessed 28 January 2023. https://www.gcsp.ch/the-geneva-process-on-AI-Principles.; Centre for Humanitarian Dialogue. 'Code of Conduct on Artificial Intelligence in Military Systems'. Centre for Humanitarian Dialogue, 2021. https://www.hdcentre.org/wp-content/uploads/2021/08/AI-Code-of-Conduct.pdf.
[538] See broadly: Bettle, Rosie. 'Mass Media Interventions: Shallow Investigation'. Founders Pledge, 17 November 2022. https://docs.google.com/document/d/1SZ590oQkLYFQtU82kdtP9WBajpqcxVTVu302lYg_Eeo/edit?
[539] For a comparative analysis of the role of para-scientific media in shaping public perceptions and policy courses in a different technological domain, that of nanotechnology, see: Kaplan, Sarah, and Joanna Radin. 'Bounding an Emerging Technology: Para-Scientific Media and the Drexler-Smalley Debate about Nanotechnology'. *Social Studies of Science* 41, no. 4 (2011): 457–85. https://www.jstor.org/stable/41301944
[540] E.g., Neal Stephenson, Cixin Liu, or many others, depending on the intended audiences.

A "lever (of governance)" can be defined as "a tool or intervention that can be used by *key actors* to shape or affect (1) the primary outcome of advanced AI development; (2) key *strategic parameters* of advanced AI governance; (3) other *key actors*' choices or *key decisions*."[541]

Research in this field includes analysis of different types of tools (key levers or interventions) available to different actors to shape advanced AI development and use.[542]

2.1. AI developer levers

Developer (intra-lab)-level levers:[543]

→ Levers for adequate AI model evaluation and technical safety testing:[544] decoding; limiting systems, adversarial training; throughout-lifecycle test, evaluation, validation, and verification (TEVV) policies;[545] internal model safety evaluations;[546] and risk assessments;[547]

[541] For discussion of these terms, see also Maas, Matthijs, 'Concepts in Advanced AI Governance: A Literature Review of Key Terms and Definitions.' *Institute for Law & AI*. AI Foundations Report 3. (October 2023). https://www.legalpriorities.org/research/advanced-ai-gov-concepts

[542] See also: Veale, Michael, Kira Matus, and Robert Gorwa. 'AI and Global Governance: Modalities, Rationales, Tensions', 2023. https://discovery.ucl.ac.uk/id/eprint/10171121/1/Veale%20Matus%20Gorwa%202023.pdf. (reviewing, and critiquing, various "modalities"—ethical codes and councils, industry governance, contracts and licensing, standards, international agreements, and converging and extraterritorial domestic regulation). See also Roberts, Huw, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. 'Global AI Governance: Barriers and Pathways Forward'. SSRN Scholarly Paper. Rochester, NY, 29 September 2023. https://doi.org/10.2139/ssrn.4588040.; A shorter mapping of governance levers and tools (called "catalysts") is provided in: Hobbhahn, Marius, Max Räuker, Yannick Mühlhäuser, Jasper Götting, and Simon Grimm. 'What Success Looks Like'. Effective Altruism Forum, 28 June 2022. https://forum.effectivealtruism.org/posts/AuRBKFnjABa6c6GzC/what-success-looks-like. (distinguishing between: "*Governance*: domestic laws, international treaties, safety regulations, whistleblower protection, auditing firms, compute governance and contingency plans; *Technical*: Red teaming, benchmarks, fire alarms, forecasting and information security; *Societal*: Norms in AI, widespread publicity, expert publicity and field-building").

[543] See also the framework in: Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, et al. 'Model Evaluation for Extreme Risks'. arXiv, 24 May 2023. https://doi.org/10.48550/arXiv.2305.15324. For other overviews, see also: Schuett, Jonas, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. 'Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion'. arXiv, 11 May 2023. https://doi.org/10.48550/arXiv.2305.07153. AI Impacts. 'Affordances for AI Labs'. AI Impacts Wiki, 25 January 2023. https://wiki.aiimpacts.org/doku.php?id=responses_to_ai:affordances:lab_affordances.; Karnofsky, Holden. 'What AI Companies Can Do Today to Help with the Most Important Century'. Cold Takes, 20 February 2023. https://www.cold-takes.com/what-ai-companies-can-do-today-to-help-with-the-most-important-century/ (drawing a distinction between interventions that support alignment research, strong security, standards and monitoring, and successful, careful AI projects).

[544] Karnofsky, Holden. 'How Might We Align Transformative AI If It's Developed Very Soon?' EA Forum, 29 August 2022. https://forum.effectivealtruism.org/posts/sW6RggfddDrcmM6Aw/how-might-we-align-transformative-ai-if-it-s-developed-very.

[545] Ashmore, Rob, Radu Calinescu, and Colin Paterson. 'Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges'. arXiv, 10 May 2019. https://doi.org/10.48550/arXiv.1905.04223.

[546] ARC Evals. 'Update on ARC's Recent Eval Efforts', 17 March 2023. https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.

[547] Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, et al. 'Model Evaluation for Extreme Risks'. arXiv, 24 May 2023. https://doi.org/10.48550/arXiv.2305.15324.

→ Levers for safe risk management in AI development process: Responsible Scaling Policies (RSPs),[548] the Three Lines of Defense (3LoD) model,[549] organizational and operational criteria for adequately safe development,[550] and "defense in depth" risk management procedures;[551]

→ Levers to ensure cautious overall decision-making: ethics and oversight boards;[552] corporate governance policies that support and enable cautious decision-making,[553] such as establishing an internal audit team;[554] and/or incorporating as a Public Benefit Corporation to allow the board of directors to balance stockholders' pecuniary interests against the corporation's social mission;

→ Levers to ensure operational security: information security best practices[555] and structured access mechanisms[556] at the level of cloud-based AI service interfaces;

→ Policies for responsibly sharing safety-relevant information: information-providing policies to increase legibility and compliance: model cards;[557]

[548] ARC Evals. 'Responsible Scaling Policies (RSPs)'. ARC Evals, 26 September 2023. https://evals.alignment.org/blog/2023-09-26-rsp/.; Anthropic. 'Anthropic's Responsible Scaling Policy, Version 1.0', 19 September 2023. http://anthropic.com/responsible-scaling-policy

[549] Schuett, Jonas. 'Three Lines of Defense against Risks from AI'. arXiv, 16 December 2022. https://doi.org/10.48550/arXiv.2212.08364.

[550] Yudkowsky, Eliezer. 'Six Dimensions of Operational Adequacy in AGI Projects'. Machine Intelligence Research Institute, 8 June 2022. https://intelligence.org/2022/06/07/six-dimensions-of-operational-adequacy-in-agi-projects/.

[551] Muelhauser, Luke. 'A Personal Take on Longtermist AI Governance'. EA Forum, 16 July 2021. https://forum.effectivealtruism.org/posts/M2SBwctwC6vBqAmZW/a-personal-take-on-longtermist-ai-governance. (footnote 19).

[552] Schuett, Jonas, Anka Reuel, and Alexis Carlier. 'How to Design an AI Ethics Board'. arXiv, 14 April 2023. https://doi.org/10.48550/arXiv.2304.07249.; for broader work on oversight boards at Meta, see: Wong, David, and Luciano Floridi. 'Meta's Oversight Board: A Review and Critical Assessment'. SSRN Scholarly Paper. Rochester, NY, 22 October 2022. https://papers.ssrn.com/abstract=4255817.; Helfer, Laurence R., and Molly K. Land. 'The Facebook Oversight Board's Human Rights Future'. SSRN Scholarly Paper. Rochester, NY, 22 August 2022. https://doi.org/10.2139/ssrn.4197107.; Kulick, Andreas. 'Corporations as Interpreters and Adjudicators of International Human Rights Norms – Meta's Oversight Board and Beyond'. SSRN Scholarly Paper. Rochester, NY, 22 September 2022. https://papers.ssrn.com/abstract=4226521.

[553] Cihon, Peter, Jonas Schuett, and Seth D. Baum. 'Corporate Governance of Artificial Intelligence in the Public Interest'. *Information* 12, no. 7 (July 2021): 275. https://doi.org/10.3390/info12070275.

[554] Schuett, Jonas. 'AGI Labs Need an Internal Audit Function'. arXiv, 26 May 2023. https://doi.org/10.48550/arXiv.2305.17038.

[555] Zabel, Claire, and Luke Muehlhauser. 'Information Security Careers for GCR Reduction'. *Effective Altruism Forum* (blog), 2019. https://forum.effectivealtruism.org/posts/ZJiCfwTy5dC4CoxqA/information-security-careers-for-gcr-reduction.; Ladish, Jeffrey. 'Information Security Considerations for AI and the Long Term Future'. EA Forum, 2 May 2022. https://forum.effectivealtruism.org/posts/WqQDCCLWbYfFRwubf/information-security-considerations-for-ai-and-the-long-term.

[556] Shevlane, Toby. 'Structured Access: An Emerging Paradigm for Safe AI Deployment'. In *The Oxford Handbook of AI Governance*, by Toby Shevlane, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang. Oxford University Press, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.39.

[557] See generally: Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 'Model Cards for Model Reporting'. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–29, 2019. https://doi.org/10.1145/3287560.3287596. And for an application, see: OpenAI. 'GPT-4 System Card'. OpenAI, 14 March 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

→ Policies to ensure organization can pace and/or pause capability research:[558] Board authority to pause research and channels to invite external AI scientists to review alignment of systems.[559]

Developer external (unilateral) levers:

→ Use of contracts and licensing to attempt to limit uses of AI and its outputs (e.g., the Responsible AI Licenses (RAIL) initiative);[560]

→ Voluntary safety commitments;[561]

→ Norm entrepreneurship (i.e., lobbying, public statements, or initiatives that signal public concern and/or dissatisfaction with an existing state of affairs, potentially alerting others to the existence of a shared complaint and facilitating potential "norm cascades" towards new expectations or collective solutions).[562]

2.2. AI industry & academia levers

Industry-level (coordinated inter-lab) levers:

→ Self-regulation;[563]
→ Codes of conduct;
→ AI ethics principles;[564]
→ Professional norms;[565]

---

[558] I thank Zach Stein-Perlman for suggesting this. For a discussion of some of the challenges involved with maintaining such policies, see: Raemon. '"Carefully Bootstrapped Alignment" Is Organizationally Hard'. LessWrong, 17 March 2023. https://www.lesswrong.com/posts/thkAtqoQwN6DtaiGT/carefully-bootstrapped-alignment-is-organizationally-hard.

[559] *The Promise of AI with Demis Hassabis - DeepMind: The Podcast (S2, Ep9)*, 2022. https://www.youtube.com/watch?v=GdeY-MrXD74.

[560] Veale, Michael, Kira Matus, and Robert Gorwa. 'AI and Global Governance: Modalities, Rationales, Tensions', 2023. https://discovery.ucl.ac.uk/id/eprint/10171121/1/Veale%20Matus%20Gorwa%202023.pdf. Pg. 8-9.

[561] Han, The Anh, Tom Lenaerts, Francisco C. Santos, and Luis Moniz Pereira. 'Voluntary Safety Commitments Provide an Escape from Over-Regulation in AI Development'. *ArXiv:2104.03741 [Nlin]*, 8 April 2021. http://arxiv.org/abs/2104.03741.

[562] The term derives from: Sunstein, Cass R. 'Social Norms and Social Roles'. *Columbia Law Review* 96, no. 4 (May 1996): 903. https://doi.org/10.2307/1123430. pg. 1996 (defining "norm entrepreneurs" as: "Political actors [who] might be able to exploit private dissatisfaction with existing norms in order to bring about large-scale social change [...] norm entrepreneurs can alert people to the existence of a shared complaint and can suggest a collective solution. [...] Thus political actors, whether public or private, can exploit widespread dissatisfaction with existing norms by (a) signaling their own commitment to change, (b) creating coalitions, (c) making defiance of the norms seem or be less costly, and (d) making compliance with new norms seem or be more beneficial.").
The norm entrepreneurship framework has been applied to many other domains, such as internet governance: Hurel, Louise Marie, and Luisa Cruz Lobato. 'Unpacking Cyber Norms: Private Companies as Norm Entrepreneurs'. *Journal of Cyber Policy* 3, no. 1 (2 January 2018): 61–76. https://doi.org/10.1080/23738871.2018.1467942.; Radu, Roxana, Matthias C. Kettemann, Trisha Meyer, and Jamal Shahin. 'Normfare: Norm Entrepreneurship in Internet Governance'. *Telecommunications Policy*, Norm entrepreneurship in Internet Governance, 45, no. 6 (1 July 2021): 102148. https://doi.org/10.1016/j.telpol.2021.102148.; multilateral arms control: Müller, Harald, and Carmen Wunderlich. *Norm Dynamics in Multilateral Arms Control: Interests, Conflicts, and Justice*. University of Georgia Press, 2013.; and others.

[563] O'Keefe, Cullen. 'Antitrust-Compliant AI Industry Self-Regulation'. LAWAI WORKING PAPER SERIES. Rochester, NY, 30 September 2021. https://doi.org/10.2139/ssrn.3933677.

[564] Schiff, D., J. Borenstein, J. Biddle, and K. Laas. 'AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection'. *IEEE Transactions on Technology and Society*, 2021, 1–1. https://doi.org/10.1109/TTS.2021.3052127.

[565] Gasser, Urs, and Carolyn Schmitt. 'The Role of Professional Norms in the Governance of Artificial Intelligence'. In *The Oxford Handbook of AI Ethics*, edited by M Dubber and F. Pasquale, 34. Oxford University Press, 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3378267

- → AI ethics advisory committees;[566]
- → Incident databases;[567]
- → Institutional, software, and hardware mechanisms for enabling developers to make verifiable claims;[568]
- → Bug bounties;[569]
- → Evaluation-based coordinated pauses;[570]
- → Other inter-lab cooperation mechanisms:[571]
  - → Assist Clause;[572]
  - → Windfall Clause;[573]
  - → Mutual monitoring agreements (red-teaming, incident-sharing, compute accounting, and seconding engineers);
  - → Communications and heads-up;
  - → Third-party auditing;
  - → Bias and safety bounties;
  - → Secure compute enclaves;
  - → Standard benchmarks & audit trails;
  - → Publication norms.[574]

Third-party industry actors levers:

- → Publication reviews;[575]

[566] Newman, Jessica Cussins. 'Decision Points in AI Governance'. Berkeley, CA: Center for Long-Term Cybersecurity, 5 May 2020. https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf.

[567] McGregor, Sean. 'Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database'. arXiv, 17 November 2020. https://doi.org/10.48550/arXiv.2011.08512. See also: Lupo, Giampiero. 'Risky Artificial Intelligence: The Role of Incidents in the Path to AI Regulation'. *Law, Technology and Humans* 5, no. 1 (30 May 2023): 133–52. https://doi.org/10.5204/lthj.2682.

[568] Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims'. *ArXiv:2004.07213 [Cs]*, 15 April 2020. http://arxiv.org/abs/2004.07213.

[569] See generally; Kenway, Josh, Camille François, Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 'Bug Bounties for Algorithmic Harms: Lessons from Cybersecurity Vulnerability Disclosure for Algorithmic Harms Discovery, Disclosure, and Redress'. Algorithmic Justice League, January 2022. https://drive.google.com/file/d/1f4hVwQNiwp13zy62wUhwIg84lOq0ciG_/view?.

[570] Alaga, Jide, and Jonas Schuett. 'Coordinated Pausing: An Evaluation-Based Coordination Scheme for Frontier AI Developers'. arXiv, 30 September 2023. https://doi.org/10.48550/arXiv.2310.00374.

[571] Askell, Amanda, Miles Brundage, and Gillian Hadfield. 'The Role of Cooperation in Responsible AI Development'. arXiv, 10 July 2019. http://arxiv.org/abs/1907.04534.

[572] See notably: OpenAI. 'OpenAI Charter'. OpenAI Blog, 9 April 2018. https://openai.com/charter.

[573] O'Keefe, Cullen, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung, and Allan Dafoe. 'The Windfall Clause: Distributing the Benefits of AI for the Common Good'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 327–31. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375842; Bova, Paolo, Jonas Emanuel Müller, and Benjamin Harack. 'Safe Transformative AI via a Windfall Clause'. *ArXiv:2108.09404 [Cs]*, 28 August 2021. http://arxiv.org/abs/2108.09404; for legal analysis, see also: Bridge, John. 'Towards a Worldwide, Watertight Windfall Clause'. EA Forum, 7 April 2022. https://forum.effectivealtruism.org/s/68dCXfuvykT3RmYy4.

[574] See list of policies enumerated in: Hua, Shin-Shin, and Haydn Belfield. 'AI & Antitrust: Reconciling Tensions Between Competition Law and Cooperative AI Development'. *Yale Journal of Law and Technology* 23 (Spring 2021): 127. (Appendix A).

[575] Wang, Jasmine. 'What the AI Community Can Learn From Sneezing Ferrets and a Mutant Virus Debate'. *Partnership on AI* (blog), 9 December 2020. https://medium.com/partnership-on-ai/lessons-for-the-ai-community-from-the-h5n1-controversy-32432438a82e.

→ Certification schemes;[576]
→ Auditing schemas.[577]

Scientific community levers:

→ Institutional Review Boards (IRBs);[578]
→ Conference or journal pre-publication impact assessment requirements;[579] academic conference practices;[580]
→ Publication and model sharing and release norms;[581]
→ Benchmarks;[582]

[576] Cihon, Peter, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum. 'AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries'. *IEEE Transactions on Technology and Society* 2, no. 4 (2021): 200–209. https://doi.org/10.1109/TTS.2021.3077595.; Winter, Philip Matthias, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, and Bernhard Nessler. 'Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications'. *ArXiv:2103.16910 [Cs, Stat]*, 31 March 2021. http://arxiv.org/abs/2103.16910.

[577] Mökander, Jakob. 'Auditing of AI: Legal, Ethical and Technical Approaches'. *Digital Society* 2, no. 3 (8 November 2023): 49. https://doi.org/10.1007/s44206-023-00074-y.; Avin, Shahar, Haydn Belfield, Miles Brundage, Gretchen Krueger, Jasmine Wang, Adrian Weller, Markus Anderljung, et al. 'Filling Gaps in Trustworthy Development of AI'. *Science (New York, N.Y.)* 374, no. 6573 (10 December 2021): 1327–29. https://doi.org/10.1126/science.abi7176.; Mökander, Jakob, Maria Axente, Federico Casolari, and Luciano Floridi. 'Conformity Assessments and Post-Market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation'. *Minds and Machines* 32, no. 2 (1 June 2022): 241–68. https://doi.org/10.1007/s11023-021-09577-4.; Raji, Inioluwa Deborah. 'From Algorithmic Audits to Actual Accountability: Overcoming Practical Roadblocks on the Path to Meaningful Audit Interventions for AI Governance'. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 5. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3514094.3539566.

[578] See Blackman, Reid. 'If Your Company Uses AI, It Needs an Institutional Review Board'. *Harvard Business Review*, 1 April 2021. https://hbr.org/2021/04/if-your-company-uses-ai-it-needs-an-institutional-review-board.; on the history of IRBs generally, see: Stark, Laura. *Behind Closed Doors: IRBs and the Making of Ethical Research*. Morality and Society Series. Chicago, IL: University of Chicago Press, 2012. https://press.uchicago.edu/ucp/books/book/chicago/B/bo12182576.html.

[579] Prunkl, Carina E. A., Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 'Institutionalizing Ethics in AI through Broader Impact Requirements'. *Nature Machine Intelligence* 3, no. 2 (February 2021): 104–10. https://doi.org/10.1038/s42256-021-00298-y.

[580] CIFAR. 'A Culture of Ethical AI: Report'. CIFAR, Partnership on AI, July 2022. https://partnershiponai.org//wp-content/uploads/dlm_uploads/2022/08/CIFAR-AI-Insights-EN-AM-220803-1.pdf.

[581] Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, et al. 'Release Strategies and the Social Impacts of Language Models'. *ArXiv:1908.09203 [Cs]*, 12 November 2019. http://arxiv.org/abs/1908.09203.; Ovadya, Aviv, and Jess Whittlestone. 'Reducing Malicious Use of Synthetic Media Research: Considerations and Potential Release Practices for Machine Learning'. arXiv, 28 July 2019. https://doi.org/10.48550/arXiv.1907.11274. See also Partnership on AI. 'Managing the Risks of AI Research: Six Recommendations for Responsible Publication'. Accessed 14 October 2022. https://partnershiponai.org/paper/responsible-publication-recommendations/.; Shevlane, Toby. 'The Artefacts of Intelligence: Governing Scientists' Contribution to AI Proliferation'. University of Oxford, 2022. https://www.governance.ai/research-paper/the-artefacts-of-intelligence-governing-scientists-contribution-to-ai-proliferation. And for a recent review: Wasil, Akash R, Charlotte Siegmann, Carson Ezell, and Aris Richardson. 'Publication Policies and Model-Sharing Decisions: A Literature Review and Recommendations for AI Labs', 2023. https://static1.squarespace.com/static/6276a63ecf564172c125f58e/t/641cbc1d84814a4d0f3e1788/1679604766050/WasilEzellRichardsonSiegmann+%2810%29.pdf.

[582] Duan, Isabella. 'Race to the Top: Rethink Benchmark-Making for Safe AI Development', 3 December 2022. https://isaduan.github.io/isabelladuan.github.io/posts/first/.; on the role of benchmarks in steering AI development, see also: Dehghani, Mostafa, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 'The Benchmark Lottery'. arXiv, 14 July 2021. https://doi.org/10.48550/arXiv.2107.07002.

→ Differential technological development (innovation prizes);[583]
→ (Temporary) moratoria.[584]

2.3. Compute supply chain industry levers

Global compute industry-level levers:[585]

→ Stock-and-flow accounting;
→ Operating licenses;
→ Supply chain chokepoints;[586]
→ Inspections
→ Passive architectural on-chip constraints (e.g., performance caps)
→ Active architectural on-chip constraints (e.g., shutdown mechanisms)

2.4. Governmental levers

We can distinguish between general governmental levers and the specific levers available to particular key states.

General governmental levers[587]

Legislatures' levers:[588]

→ Create new AI-specific regimes, such as:

---

[583] Sandbrink, Jonas, Hamish Hobbs, Jacob Swett, Allan Dafoe, and Anders Sandberg. 'Differential Technology Development: A Responsible Innovation Principle for Navigating Technology Risks'. SSRN Scholarly Paper. Rochester, NY, 8 September 2022. https://papers.ssrn.com/abstract=4213670. (on the general principle, not specifically focused on AI); However, for a skeptical take on the efficacy of innovation prizes, see: Howes, Anton. 'Why Innovation Prizes Fail'. *Works in Progress* (blog), 21 April 2022. https://www.worksinprogress.co/issue/why-innovation-prizes-fail/.

[584] Vöneky, Silja. 'Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks'. In *Human Rights, Democracy, and Legitimacy in a World of Disorder*, edited by Silja Vöneky and Gerald Neuman, 139–62. Cambridge University Press, 2018. https://papers.ssrn.com/abstract=3363552. (discussing scientific moratoria generally).

[585] Heim, Lennart. 'Transformative AI and Compute'. AI Alignment Forum, 2021. https://www.alignmentforum.org/s/bJi3hd8E8qjBeHz9Z. See also Vipra, Jai, and Sarah Myers West. 'Computational Power and AI'. AI Now Institute, 27 September 2023. https://ainowinstitute.org/publication/policy/compute-and-ai.

[586] Barbe, Andre, and Will Hunt. 'Preserving the Chokepoints: Reducing the Risks of Offshoring Among U.S. Semiconductor Manufacturing Equipment Firms'. Center for Security and Emerging Technology, May 2022. https://cset.georgetown.edu/publication/preserving-the-chokepoints/.

[587] See also: Scherer, Matthew U. 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies'. *Harvard Journal of Law & Technology*, no. 2 (Spring 2016). http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf. (reviewing legislatures, expert agencies, and the common law tort system); see also Winter, Christoph, Jonas Schuett, Eric Martínez, Suzanne Van Arsdale, Renan Araújo, Nick Hollman, Jeff Sebo, Andrew Stawasz, Cullen O'Keefe, and Giuliana Rotola. 'Legal Priorities Research: A Research Agenda'. Legal Priorities Project, January 2021. https://www.legalpriorities.org/research_agenda.pdf. See also: AI Impacts. 'Affordances for States'. AI Impacts Wiki, 25 January 2023. https://wiki.aiimpacts.org/doku.php?id=responses_to_ai:affordances:state_affordances. And broadly: Karnofsky, Holden. 'How Major Governments Can Help with the Most Important Century', 24 February 2023. https://forum.effectivealtruism.org/posts/ruJnXtdDS7XiiwzSP/how-major-governments-can-help-with-the-most-important.

[588] There are various collections that discuss the regulation of AI on the basis of extant bodies of law, though these focus primarily on the regulation of algorithms that exist today, rather than of more capable or transformative AI systems. See e.g. Barfield, Woodrow, and Ugo Pagallo, eds. *Research Handbook on the Law of Artificial Intelligence*. Cheltenham, UK: Edward Elgar Publishing, 2018. https://www.elgaronline.com/view/edcoll/9781786439048/9781786439048.xml.; Wischmeyer, Thomas, and Timo Rademacher, eds. *Regulating Artificial Intelligence*. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-32361-5; DiMatteo, Larry A., Cristina Poncibò, and Michel Cannarsa, eds. *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*. Cambridge Law Handbooks. Cambridge: Cambridge University Press, 2022. https://doi.org/10.1017/9781009072168.

- → Horizontal risk regulation;[589]
- → Industry-specific risk regulatory regimes;
- → Permitting, licensing, and market gatekeeping regimes;[590]
- → Bans or moratoria;
- → Know-Your-Customer schemes.[591]

→ Amend laws to extend or apply existing regulations to AI:[592]

- → Domain/industry-specific risk regulations;
- → Competition/antitrust law,[593] including doctrines around merger control, abuse of dominance, cartels, and collusion; agreements on hardware security; and state aid;
- → Liability law;[594]
- → Insurance law;[595]
- → Contract law;[596]
- → IP law;[597]
- → Copyright law (amongst others through its impact on data scraping practices);[598]
- → Criminal law;[599]

[589] Petit, Nicolas, and Jerome De Cooman. 'Models of Law and Regulation for AI'. EUI Working Paper RSCAS 2020/63. Social Science Research Network, 1 October 2020. https://doi.org/10.2139/ssrn.3706771.; see also Maas, Matthijs M. 'Aligning AI Regulation to Sociotechnical Change'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang. Oxford University Press, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.22.

[590] Higgins, Brian W. 'Legal Elements of an AI Regulatory Permit Program'. In *The Oxford Handbook of AI Governance*, by Brian Wm. Higgins, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang. Oxford University Press, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.18. Malgieri, Gianclaudio, and Frank Pasquale. 'Licensing High-Risk Artificial Intelligence: Toward Ex Ante Justification for a Disruptive Technology'. *Computer Law & Security Review* 52 (1 April 2024): 105899. https://doi.org/10.1016/j.clsr.2023.105899.

[591] Egan, Janet, and Lennart Heim. 'Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers'. arXiv, 20 October 2023. http://arxiv.org/abs/2310.13625.

[592] For a distinction of regulatory responses between "drawing analogies," "extending existing law," "creating new law," and "reassessing the regulatory regime," see: Crootof, Rebecca, and B. J. Ard. 'Structuring Techlaw'. *Harvard Journal of Law & Technology* 34, no. 2 (2021): 347–417. https://jolt.law.harvard.edu/assets/articlePDFs/v34/1.-Crootof-Ard-Structuring-Techlaw.pdf

[593] Hua, Shin-Shin, and Haydn Belfield. 'AI & Antitrust: Reconciling Tensions Between Competition Law and Cooperative AI Development'. *Yale Journal of Law and Technology* 23 (Spring 2021): 127. https://yjolt.org/ai-antitrust-reconciling-tensions-between-competition-law-and-cooperative-ai-development

[594] White, Trevor N., and Seth D. Baum. 'Liability For Present And Future Robotics Technology'. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, 2017, 5.; Erdélyi, Olivia J., and Gábor Erdélyi. 'The AI Liability Puzzle and a Fund-Based Work-Around'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 50–56. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375806.

[595] Lior, Anat. 'Insuring AI: The Role of Insurance in Artificial Intelligence Regulation'. *Harvard Journal of Law & Technology* 35, no. 2 (2022): 64. https://jolt.law.harvard.edu/assets/articlePDFs/v35/2.-Lior-Insuring-AI.pdf

[596] See also: Linarelli, John. 'Artificial General Intelligence and Contract'. *Uniform Law Review* 24, no. 2 (1 June 2019): 330–47. https://doi.org/10.1093/ulr/unz015.

[597] Calvin, Nathan, and Jade Leung. 'Who Owns Artificial Intelligence? A Preliminary Analysis of Corporate Intellectual Property Strategies and Why They Matter'. Oxford: Center for the Governance of AI, Future of Humanity Institute, February 2020. https://www.fhi.ox.ac.uk/wp-content/uploads/Patents_-FHI-Working-Paper-Final-.pdf.; and previously Koepsell, David. 'Can the Singularity Be Patented? (And Other IP Conundrums for Converging Technologies)'. In *The Technological Singularity: Managing the Journey*, edited by Victor Callaghan, James Miller, Roman Yampolskiy, and Stuart Armstrong, 181–91. The Frontiers Collection. Berlin, Heidelberg: Springer, 2017. https://doi.org/10.1007/978-3-662-54033-6_10.

[598] See also Vincent, James. 'The Lawsuit That Could Rewrite the Rules of AI Copyright'. The Verge, 8 November 2022. https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data.

[599] See generally King, Thomas C., Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi. 'Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions'. *Science and Engineering Ethics*, 14 February 2019. https://doi.org/10.1007/s11948-018-00081-0.

- → Privacy and data protection law (amongst others through its impact on data scraping practices);
- → Public procurement law and procurement processes.[600]

Executive levers:

- → Executive orders;
- → Foreign investment restrictions;
- → AI R&D funding strategies;[601]
- → Nationalization of firms;
- → Certification schemes;
- → Various tools of "differential technology development":[602] policies for preferential advancement of safer AI architectures (funding and direct development programs, government prizes, advanced market commitments, regulatory requirements, and tax incentives)[603] and policies for slowing down research lines towards dangerous AI architectures (moratoria, bans, defunding, divestment, and/or "stage-gating" review processes);[604]
- → Foreign policy decisions, such as initiating multilateral treaty negotiations.

Judiciaries' levers:

- → Judicial decisions handed down on cases involving AI that extend or apply existing doctrines to AI, shaping economic incentives and setting precedent for regulatory treatment of advanced AI, such as the US Supreme Court ruling on *Gonzalez v. Google*, which has implications for whether algorithmic recommendations will receive full Section 230 protections;[605]
- → Judicial review, especially of drastic executive actions taken in response to AI risk scenarios;[606]
- → Judicial policymaking, through discretion in evaluating proportionality or balancing tests.[607]

Expert agencies' levers:

---

[600] Belfield, Haydn, Amritha Jayanti, and Shahar Avin. 'Written Evidence - Defence Industrial Policy: Procurement and Prosperity', 2020. https://committees.parliament.uk/writtenevidence/4785/default/.; See generally: Dor, Lavi M. Ben, and Cary Coglianese. 'Procurement as AI Governance'. *IEEE Transactions on Technology and Society*, 2021, 1–1. https://doi.org/10.1109/TTS.2021.3111764.

[601] Dawson, Gregory, Kevin Desouza, and James Denford. 'Understanding Artificial Intelligence Spending by the U.S. Federal Government'. *Brookings* (blog), 22 September 2022. https://www.brookings.edu/blog/techtank/2022/09/22/understanding-artificial-intelligence-spending-by-the-u-s-federal-government/.

[602] Sandbrink, Jonas, Hamish Hobbs, Jacob Swett, Allan Dafoe, and Anders Sandberg. 'Differential Technology Development: A Responsible Innovation Principle for Navigating Technology Risks'. SSRN Scholarly Paper. Rochester, NY, 8 September 2022. https://papers.ssrn.com/abstract=4213670.

[603] Ibid.

[604] Ibid.

[605] Perault, Matt. 'Section 230 Won't Protect ChatGPT'. Lawfare, 23 February 2023. https://www.lawfareblog.com/section-230-wont-protect-chatgpt.; Robertson, Adi. 'The Supreme Court Could Be About to Decide the Legal Fate of AI Search'. The Verge, 16 February 2023. https://www.theverge.com/2023/2/16/23591290/supreme-court-section-230-gonzalez-google-bard-bing-ai-search-algorithms. And see generally: Kosseff, Jeff. 'A User's Guide to Section 230, and a Legislator's Guide to Amending It (or Not)'. *Berkeley Technology Law Journal* 37, no. 2 (2022). https://papers.ssrn.com/abstract=3905347.

[606] There has been little direct work on applying this to AI; however, for discussions of this in the context of COVID responses, see: Ginsburg, Tom, and Mila Versteeg. 'The Bound Executive: Emergency Powers during the Pandemic'. *International Journal of Constitutional Law* 19, no. 5 (1 December 2021): 1498–1535. https://doi.org/10.1093/icon/moab059. I thank Christoph Winter for this suggestion.

[607] For the use of courts in other domains, see: Martinsen, Dorte Sindbjerg. 'Judicial Policy-Making and Europeanization: The Proportionality of National Control and Administrative Discretion'. *Journal of European Public Policy* 18, no. 7 (1 October 2011): 944–61. https://doi.org/10.1080/13501763.2011.599962.; I thank Christoph Winter for this suggestion.

- → A mix of features of other actors, from setting policies to adjudicating disputes to enforcing decisions;[608]
- → Create or propose soft law.[609]

Ancillary institutions:

- → Improved monitoring infrastructures;[610]
- → Provide services in terms of training, insurance, procurement, identification, archiving, etc.[611]

Foreign Ministries/State Department:

- → Set activities and issue agendas in global AI governance institutions;
- → Bypass or challenge existing institutions by engaging in "competitive regime creation,"[612] "forum shopping,"[613] or the strategic creation of treaty conflicts;[614]
- → Initiate multilateral treaty negotiations;
- → Advice policymakers about the existence and meaning of international law and which obligations these impose;[615]
- → Conduct state behavior around AI issues (in terms of state policy, and through discussion of AI issues in national legislation, diplomatic correspondence, etc.) in such a way as to contribute to the establishment of binding customary international law (CIL).[616]

---

[608] Scherer, Matthew U. 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies'. *Harvard Journal of Law & Technology*, no. 2 (Spring 2016). http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf. Pg. 382.

[609] Villasenor, John. 'Soft Law as a Complement to AI Regulation'. *Brookings* (blog), 31 July 2020. https://www.brookings.edu/research/soft-law-as-a-complement-to-ai-regulation/. See also Marchant, Gary E., and Carlos Ignacio Gutierrez. 'Indirect Enforcement of Artificial Intelligence "Soft Law"'. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 15 December 2020. https://doi.org/10.2139/ssrn.3749776.; Gutierrez, Carlos Ignacio, Gary E. Marchant, and Katina Michael. 'Effective and Trustworthy Implementation of AI Soft Law Governance'. *IEEE Transactions on Technology and Society* 2, no. 4 (December 2021): 168–70. https://doi.org/10.1109/TTS.2021.3121959.

[610] Whittlestone, Jess, and Jack Clark. 'Why and How Governments Should Monitor AI Development'. *ArXiv:2108.12427 [Cs]*, 31 August 2021. http://arxiv.org/abs/2108.12427.; also Clark, Jack. 'Technical Observatories for Better AI Governance'. In *The Oxford Handbook of AI Governance*, edited by Valerie Hudson and Justin Bullock. Oxford Univ. Press, [forthcoming].

[611] Hudson, Valerie M. 'Standing Up a Regulatory Ecosystem for Governing AI Decision-Making: Principles and Components'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press. Accessed 21 October 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.17.

[612] See generally Morse, Julia C., and Robert O. Keohane. 'Contested Multilateralism'. *The Review of International Organizations* 9, no. 4 (1 December 2014): 385–412. https://doi.org/10.1007/s11558-014-9188-2.

[613] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Should Artificial Intelligence Governance Be Centralised?: Design Lessons from History'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–34. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375857.; and see generally: Fehl, Caroline. 'Forum Shopping from above and below: Power Shifts and Institutional Choice in a Stratified International Society', 36. Munich, 2016.

[614] See generally: Ranganathan, Surabhi. *Strategically Created Treaty Conflicts and the Politics of International Law*. Cambridge Studies in International and Comparative Law. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781107338005.

[615] Deeks, Ashley. 'High-Tech International Law'. *George Washington Law Review* 88 (2020): 575–653. See pg. 590-591. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531976

[616] For in-depth discussion of the role of CIL, see: Hakimi, Monica. 'Making Sense of Customary International Law'. *Michigan Law Review* 118 (16 June 2020). https://papers.ssrn.com/abstract=3627905. ; for an analysis of the merits and roles of CIL, see: Helfer, Laurence R, and Ingrid B Wuerth. 'Customary International Law: An Instrument Choice Perspective'. *Michigan Journal of International Law* 37 (2016): 563. https://repository.law.umich.edu/mjil/vol37/iss4/1/ ; Crootof has argued that changing state practice may even modify established treaty law; Crootof, Rebecca. 'Change Without Consent: How Customary International Law Modifies Treaties'. *Yale Journal of International Law* 41, no. 2 (2016): 65. https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1670&context=yjil

Specific key governments levers
Levers available to specific key governments:

US-specific levers:[617]

- → AI-specific regulations, such as the AI Bill of Rights;[618] Algorithmic Accountability Act;[619] 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence;[620] and various currently pending federal legislative proposals for regulating generative and/or frontier AI;[621]
- → General levers,[622] such as federal R&D funding, foreign investment restrictions, export controls,[623] visa vetting, expanded visa pathways, secrecy orders, voluntary screening procedures, use of the Defense Production Act,[624] antitrust enforcement, the "Born Secret" Doctrine, nationalization of companies or compute hardware, various Presidential Emergency powers,[625] etc.

EU-specific levers:

- → AI-specific regulations, including:

[617] See also the overview in Pouget, Hadrien Pouget, Matt, and Matthew O'Shaughnessy. 'Reconciling the U.S. Approach to AI'. Carnegie Endowment for International Peace, 3 May 2023. https://carnegieendowment.org/2023/05/03/reconciling-u.s.-approach-to-ai-pub-89674.
[618] The White House. 'Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People', October 2022, 73. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf
[619] Rep. Clarke, Yvette D. [D-NY-9. 'Text - H.R.6580 - 117th Congress (2021-2022): Algorithmic Accountability Act of 2022'. Legislation, 2 April 2022. 02/04/2022. https://www.congress.gov/bill/117th-congress/house-bill/6580/text
[620] Biden, Joseph R. 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'. The White House, 30 October 2023. 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/.
[621] Lenhart, Anna. 'Roundup of Federal Legislative Proposals That Pertain to Generative AI'. Tech Policy Press, 21 April 2023. https://techpolicy.press/roundup-of-federal-legislative-proposals-that-pertain-to-generative-ai/.; for an overview see also: Matthews, Dylan. 'The AI Rules That US Policymakers Are Considering, Explained'. Vox, 1 August 2023. https://www.vox.com/future-perfect/23775650/ai-regulation-openai-gpt-anthropic-midjourney-stable.
[622] Fischer, Sophie-Charlotte, Jade Leung, Markus Anderljung, Cullen O'Keefe, Stefan Torges, Saif M. Khan, Ben Garfinkel, and Allan Dafoe. 'AI Policy Levers: A Review of the U.S. Government's Tools to Shape AI Research, Development, and Deployment'. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, March 2021. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/AI-Policy-Levers-A-Review-of-the-U.S.-Governments-tools-to-shape-AI-research-development-and-deployment-%E2%80%93-Fischer-et-al.pdf.
[623] Flynn, Carrick. 'Recommendations on Export Controls for Artificial Intelligence'. Center for Security and Emerging Technology, February 2020. https://cset.georgetown.edu/research/recommendations-on-export-controls-for-artificial-intelligence/. Leung, Jade, Sophie-Charlotte Fischer, and Allan Dafoe. 'Export Controls in the Age of AI'. War on the Rocks, 28 August 2019. https://warontherocks.com/2019/08/export-controls-in-the-age-of-ai/. For more recent work, see: Brockmann, Kolja. 'Applying Export Controls to AI: Current Coverage and Potential Future Controls'. In *Armament, Arms Control and Artificial Intelligence: The Janus-Faced Nature of Machine Learning in the Military Realm*, edited by Thomas Reinhold and Niklas Schörnig, 193–209. Studies in Peace and Security. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-031-11043-6_14.
[624] See generally Baker, James E. 'A DPA for the 21st Century'. Center for Security and Emerging Technology, April 2021. https://cset.georgetown.edu/publication/a-dpa-for-the-21st-century/.
[625] Brennan Center for Justice. 'A Guide to Emergency Powers and Their Use', February 2023. https://www.brennancenter.org/our-work/research-reports/guide-emergency-powers-and-their-use (listing up to 148 statutory powers that become available upon declaration of war and/or a national emergency).

- → The AI Act, which will have direct regulatory effects[626] but may also exert extraterritorial impact as part of a "Brussels Effect";[627]
- → Standard-setting by European Standards Organizations (ESOs);[628]
- → AI Liability Directive.[629]

China-specific levers:

- → AI-specific regulations;[630]
- → Standards;[631]
- → Activities in global AI governance institutions.[632]

UK-specific levers:[633]

- → National Security and Investment Act 2021;
- → Competition Law: 1998 Competition Act;
- → Export Control legislation;
- → Secrecy orders.

2.5. Public, civil society & media actor levers

Civil Society/activist movement levers:[634]

---

[626] Stix, Charlotte. 'The Ghost of AI Governance Past, Present, and Future: AI Governance in the European Union'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press. Accessed 21 October 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.56.

[627] Siegmann, Charlotte, and Markus Anderljung. 'The Brussels Effect and Artificial Intelligence': Centre for the Governance of AI, August 2022. https://www.governance.ai/research-paper/brussels-effect-ai. See also broadly: Dempsey, Mark, Keegan McBride, Meeri Haataja, and Joanna J. Bryson. 'Transnational Digital Governance and Its Impact on Artificial Intelligence'. In *The Oxford Handbook of AI Governance*, by Mark Dempsey, Keegan McBride, Meeri Haataja, and Joanna J. Bryson, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang. Oxford University Press, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.16.; but for a critical discussion, see: Almada, Marco, and Anca Radu. 'The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy'. SSRN Scholarly Paper. Rochester, NY, 9 June 2023. https://papers.ssrn.com/abstract=4592006.

[628] O'Keefe, Cullen, Jade Leung, and Markus Anderljung. 'How Technical Safety Standards Could Promote TAI Safety'. Effective Altruism Forum, 8 August 2022. https://forum.effectivealtruism.org/posts/zvbGXCxc5jBowCuNX/how-technical-safety-standards-could-promote-tai-safety.

[629] European Commission. 'Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive)'. European Commission, 28 September 2022. https://ec.europa.eu/info/sites/default/files/1_1_197605_prop_dir_ai_en.pdf.

[630] DigiChina. 'How Will China's Generative AI Regulations Shape the Future? A DigiChina Forum', 19 April 2023. https://digichina.stanford.edu/work/how-will-chinas-generative-ai-regulations-shape-the-future-a-digichina-forum/.

[631] Sheehan, Matt. 'China's New AI Governance Initiatives Shouldn't Be Ignored'. Carnegie Endowment for International Peace, 4 January 2022. https://carnegieendowment.org/2022/01/04/china-s-new-ai-governance-initiatives-shouldn-t-be-ignored-pub-86127.

[632] Cheng, Jing, and Jinghan Zeng. 'Shaping AI's Future? China in Global AI Governance'. *Journal of Contemporary China* 0, no. 0 (8 August 2022): 1–17. https://doi.org/10.1080/10670564.2022.2107391.

[633] The below is based on the review in: Hadshar, Rose. 'Current UK Government Levers on AI Development'. EA Forum, 10 April 2023. https://forum.effectivealtruism.org/posts/BFBf5yPLoJMGozygE/current-uk-government-levers-on-ai-development.

[634] For suggestions here, I also thank James Ozden. See also more generally Ozden, James, and Sam Glover. 'Protest Movements: How Effective Are They?' Social Change Lab, July 2022. https://www.socialchangelab.org/_files/ugd/503ba4_052959e2ee8d4924934b7efe3916981e.pdf.; see also the taxonomy in: Beer, Michael A. 'Civil Resistance Tactics in the 21st Century'. International Center on Nonviolent Conflict, 2021. https://www.vredesmuseum.nl/download/civilresistance.pdf.

- → Lab-level (internal) levers:
  - → Shareholder activism, voting out CEOs;
  - → Unions and intra-organizational advocacy, strikes, and walkouts;[635]
  - → Capacity-building of employee activism via recruitment, political education, training, and legal advice.

- → Lab-level (external) levers:
  - → Stigmatization of irresponsible practices;[636]
  - → Investigative journalism, awareness-raising of scandals and incidents, hacking and leaks, and whistleblowing;
  - → Impact litigation[637] and class-action lawsuits;[638]
  - → Public protest[639] and direct action (e.g., sit-ins).

- → Industry-level levers:
  - → Norm advocacy and lobbying;
  - → Open letters and statements;
  - → Mapping and highlighting (compliance) performance of companies; establishing metrics, indexes, and prizes; and certification schemes.[640]

- → Public-focused levers:
  - → Media content creation;[641]
  - → Boycott and divestment;
  - → Shaming of state noncompliance with international law;[642]
  - → Emotional contagion—shaping and disseminating of public emotional dynamics or responses to a crisis.[643]

- → Creating alternatives:
  - → Public interest technology research;

---

[635] Belfield, Haydn. 'Activism by the AI Community: Analysing Recent Achievements and Future Prospects'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 15–21. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375814.

[636] Baum, Seth D. 'On the Promotion of Safe and Socially Beneficial Artificial Intelligence'. *AI & SOCIETY*, 28 September 2016. https://doi.org/10.1007/s00146-016-0677-0.

[637] See generally; AI Now Institute. 'Taking Algorithms To Court'. *Medium* (blog), 24 September 2018. https://medium.com/@AINowInstitute/taking-algorithms-to-court-7b90f82ffcc9.; for an overview of (US) cases, see Ethical Tech Initiative of DC. 'AI Litigation Database'. Accessed 20 October 2022. https://blogs.gwu.edu/law-eti/ai-litigation-database/.

[638] See e.g. Vincent, James. 'The Lawsuit That Could Rewrite the Rules of AI Copyright'. The Verge, 8 November 2022. https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data.; Butterick, Matthew. 'Stable Diffusion Litigation', 13 January 2023. https://stablediffusionlitigation.com/.

[639] See generally: Hobson, Tom. 'Kill the Bill to Save The Future'. *Medium* (blog), 29 December 2021. https://medium.com/@t.hobson/kill-the-bill-to-save-the-future-e62689e02328 (discussing generally the importance of protest to existential risk mitigation).

[640] Cihon, Peter, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum. 'AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries'. *IEEE Transactions on Technology and Society* 2, no. 4 (December 2021): 200–209. https://doi.org/10.1109/TTS.2021.3077595.

[641] See for example: *Slaughterbots - If Human: Kill()*, 2021. https://www.youtube.com/watch?v=9rDo1QxI260.

[642] See generally: Dothan, Shai. 'Social Networks and the Enforcement of International Law'. SSRN Scholarly Paper. Rochester, NY, 2 May 2017. https://doi.org/10.2139/ssrn.2961715.

[643] See generally: Holthaus, Leonie. 'Feelings of (Eco-) Grief and Sorrow: Climate Activists as Emotion Entrepreneurs'. *European Journal of International Relations* 29, no. 2 (1 June 2023): 352–73. https://doi.org/10.1177/13540661221136772.

→ Creating alternative (types of) institutions[644] and new AI labs.

→ State-focused levers:
→ Monitor compliance with international law.[645]

2.6. International organizations and regime levers

International standards bodies' levers:

→ Set technical safety and reliability standards;[646]
→ Undertake "para-regulation," setting pathways for future regulation not by imposing substantive rules but rather by establishing foundational concepts or terms.[647]

International regime levers:[648]

→ Setting or shaping norms and expectations:
→ Setting, affirming, and/or clarifying states' obligations under existing international law principles;
→ Set fora and/or agenda for negotiation of new treaties or regimes in various formats, such as:
→ Broad framework conventions;[649]
→ Nonproliferation and arms control agreements;[650]
→ Export control regimes.[651]
→ Create (technical) benchmarks and focal points for decision-making by both states and non-state actors;[652]
→ Organize training and workshops with national officials.

→ Coordinating behavior; reducing uncertainty, improving trust:

---

[644] See e.g. The Collective Intelligence Project. 'Whitepaper'. The Collective Intelligence Project, 2023. https://cip.org/whitepaper.

[645] See generally: Eilstrup-Sangiovanni, Mette, and J. C. Sharman. *Vigilantes beyond Borders: NGOs as Enforcers of International Law*. *Vigilantes beyond Borders*. Princeton University Press, 2022. https://doi.org/10.1515/9780691232249.

[646] Cihon, Peter. 'Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development'. Technical Report. Oxford: Center for the Governance of AI, Future of Humanity Institute, University of Oxford, April 2019. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.; O'Keefe, Cullen, Jade Leung, and Markus Anderljung. 'How Technical Safety Standards Could Promote TAI Safety'. Effective Altruism Forum, 8 August 2022. https://forum.effectivealtruism.org/posts/zvbGXCxc5jBowCuNX/how-technical-safety-standards-could-promote-tai-safety.

[647] Villarino, José-Miguel Bello y. 'Global Standard-Setting for Artificial Intelligence: Para-Regulating International Law for AI?' *The Australian Year Book of International Law Online* 41, no. 1 (23 October 2023): 157–81. https://doi.org/10.1163/26660229-04101018.

[648] I thank José Jaime Villalobos for input and suggestions on this section.

[649] See generally: Matz-Lück, Nele. 'Framework Conventions as a Regulatory Tool'. *Goettingen Journal of International Law* 3 (2009): 439–58. https://doi.org/10.3249/1868-1581-1-3-MATZ-LUECK.

[650] Scharre, Paul, and Megan Lamberth. 'Artificial Intelligence and Arms Control'. Center for a New American Security, 12 October 2022. https://www.cnas.org/publications/reports/artificial-intelligence-and-arms-control. Maas, Matthijs M. 'How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons'. *Contemporary Security Policy* 40, no. 3 (6 February 2019): 285–311. https://doi.org/10.1080/13523260.2019.1576464.

[651] Brockmann, Kolja. 'Applying Export Controls to AI: Current Coverage and Potential Future Controls'. In *Armament, Arms Control and Artificial Intelligence: The Janus-Faced Nature of Machine Learning in the Military Realm*, edited by Thomas Reinhold and Niklas Schörnig, 193–209. Studies in Peace and Security. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-031-11043-6_14.

[652] See generally Howse, Robert, and Ruti Teitel. 'Beyond Compliance: Rethinking Why International Law Really Matters'. *Global Policy* 1, no. 2 (2010): 127–36. https://doi.org/10.1111/j.1758-5899.2010.00035.x.

- → Confidence-building measures;[653]
- → Review conferences (e.g., BWC);
- → Conferences of parties (e.g., UNFCCC);
- → Establishing information and benefit-sharing mechanisms.

→ Creating common knowledge or shared perceptions of problems; establish "fire alarms":
- → Intergovernmental scientific bodies (e.g., Intergovernmental Panel on Climate Change (IPCC) and Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES));
- → International warning systems (e.g., WHO's "public health emergency of international concern" mechanism).

→ Adjudicating and arbitrating state disagreements over application of policies, resolving tensions or crises for regimes:
- → Arbitral bodies (e.g., WTO Appellate Body);
- → Adjudicatory tribunals (e.g., ICJ);
- → Treaty bodies (e.g., Human Rights Committee);
- → Other dispute resolution mechanisms (e.g., BWC or Resolution 1540 allowing complaints to be lodged at the UNSC).

→ Establishing material constraints:
- → Supply-side material proliferation controls (e.g., stock-and-flow accounting and trade barriers);
- → Fair and equitable treatment standards in international investment law.

→ Monitoring state compliance:
- → Inspection regimes;
- → Safeguards;
- → National contributions;
- → Network of national contact points.

→ Sanctioning noncompliance:
- → Inducing direct costs through sanctions;
- → Inducing reputational costs,[654] in particular through shaming.[655]

2.7. Future, new types of institutions and levers

Novel governance institutions and innovations:

---

[653] Ruhl, Christian. 'Risks from Autonomous Weapon Systems and Military AI'. Founders Pledge, 19 May 2022. https://forum.effectivealtruism.org/posts/RKMNZn7r6cT2Yaorf/risks-from-autonomous-weapon-systems-and-military-ai. See also: Horowitz, Michael C, and Paul Scharre. 'AI and International Stability: Risks and Confidence-Building Measures'. Center for a New American Security, 12 January 2021. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.; Horowitz, Michael C., Lauren Kahn, and Casey Mahoney. 'The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?' *Orbis*, 14 September 2020. https://doi.org/10.1016/j.orbis.2020.08.003.

[654] Guzman, Andrew T. 'The Design of International Agreements'. *European Journal of International Law* 16, no. 4 (1 September 2005): 579–612. https://doi.org/10.1093/ejil/chi134. For how new technologies may affect this, see: McGregor, Lorna. 'Are New Technologies an Aid to Reputation as a Disciplinarian?' *AJIL Unbound* 113 (ed 2019): 238–41. https://doi.org/10.1017/aju.2019.54.

[655] Dothan, Shai. 'A Virtual Wall of Shame: The New Way of Imposing Reputational Sanctions on Defiant States'. *Duke Journal of Comparative and International Law* 27 (2017 2016): 141.

- → "Regulatory markets" and private regulatory authorities;[656]
- → New monitoring institutions and information markets;[657]
- → Quadratic voting and radical markets[658]
- → Blockchain smart contracts.[659]

3. Pathways to influence (on each key actor)

That is, how might concerned stakeholders ensure that key actors use their levers to shape advanced AI development in appropriate ways?

> In this context, a "pathway (to influence)" can be defined as "a tool or intervention by which other actors (that may not themselves be key actors) can affect, persuade, induce, incentivize, or require *key actors* to make certain *key decisions* around the governance of AI. This can include interventions that ensure that certain *levers of control* are (not) used, or used in particular ways."[660]

This includes research on the different pathways by which the use of these above levers might be enabled, advocated for, and implemented (i.e., the tools available to affect the decisions by key actors).

This can draw on mappings and taxonomies: "A Map to Navigate AI Governance"[661] "The Longtermist AI Governance Landscape".[662]

---

[656] Hadfield, Gillian K., and Jack Clark. 'Regulatory Markets: The Future of AI Governance'. arXiv, 25 April 2023. https://doi.org/10.48550/arXiv.2304.04914. See previously Clark, Jack, and Gillian K Hadfield. 'Regulatory Markets for AI Safety', Safe Machine Learning workshop at ICLR, 2019. 2019. https://arxiv.org/abs/2001.00078.

[657] Clark, Jack. 'Information Markets and AI Development'. In *The Oxford Handbook of AI Governance*, edited by Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, and Baobao Zhang, 0. Oxford University Press. Accessed 6 February 2023. https://doi.org/10.1093/oxfordhb/9780197579329.013.21.

[658] See generally: Posner, Eric A., and Eric Glen Weyl. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton ; Oxford: Princeton University Press, 2018.

[659] Buterin, Vitalik. 'Why Cryptoeconomics and X-Risk Researchers Should Listen to Each Other More'. Medium, 5 July 2016. https://medium.com/@VitalikButerin/why-cryptoeconomics-and-x-risk-researchers-should-listen-to-each-other-more-a2db72b3e86b.

[660] For definitions, see also Maas, Matthijs, 'Concepts in advanced AI governance: A literature review of key terms and definitions.' *Institute for Law & AI*. AI Foundations Report 3. (October 2023). https://www.legalpriorities.org/research/advanced-ai-gov-concepts

[661] Moës, Nicolas, and Caroline Jeanmaire. 'A Map to Navigate AI Governance'. EA Forum, 14 February 2022. https://forum.effectivealtruism.org/posts/tmxkRFx6HyhhvHdz4/a-map-to-navigate-ai-governance. (highlighting three major "governance pathways"—hard governance, industry-wide self-governance, and company self-governance—each with associated sub-activities; also mentions a range of additional governance pathways not mentioned on the map: military and national security governance, supply chain and trade governance, multilateral soft governance, extralegal governance, and academic governance).

[662] Clarke, Sam. 'The Longtermist AI Governance Landscape: A Basic Overview'. EA Forum, 18 January 2022. https://forum.effectivealtruism.org/posts/ydpo7LcJWhrr2GJrx/the-longtermist-ai-governance-landscape-a-basic-overview. ("sketches a spectrum of activities, spanning strategy research, tactics research, policy development work, policy advocacy work, and policy implementation work—supported by field-building work at all levels.").

3.1. Pathways to directly shaping advanced AI systems' actions through law

Directly shaping advanced AI actions through law (i.e., legal systems and norms as an anchor or lodestar for technical alignment approaches):

- → "Law-following AI";[663]
- → Encode "incomplete contracting" as a framework for AI alignment;[664]
- → Negative human rights as technical safety constraint for minimal alignment;[665]
- → Human rights norms as a benchmark for maximal alignment;[666]
- → Encode fiduciary duties towards users into AI systems;[667]
- → Mandatory on-chip controls (monitoring and remote shutdown);
- → Legal informatics approach to alignment.[668]

3.2. Pathways to shaping governmental decisions

Shaping governmental decisions around AI levers at the level of:

- → Legislatures:
  - → Advocacy within the legislative AI policymaking process.[669]
- → Executives:
  - → Serve as high-bandwidth policy advisor;[670]
  - → Provide actionable technical information;[671]

---

[663] O'Keefe, Cullen. 'Law-Following AI'. AI Alignment Forum, 4 August 2022. https://www.alignmentforum.org/s/ZytYxd523oTnBNnRT.

[664] Hadfield-Menell, Dylan, and Gillian Hadfield. 'Incomplete Contracting and AI Alignment'. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019. http://arxiv.org/abs/1804.04268.

[665] Bajgar, Ondrej, and Jan Horenovsky. 'Negative Human Rights as a Basis for Long-Term AI Safety and Regulation'. *Journal of Artificial Intelligence Research*, 2022, 30. https://arxiv.org/abs/2208.14788; see also Bajgar, Ondrej, and Jan Horenovsky. 'Narrow Rules are not Enough: Why artificial intelligence needs to understand human rights'. *Verfassungsblog* (blog), 11 August 2022. https://verfassungsblog.de/narrow-rules-are-not-enough/.

[666] Gabriel, Iason. 'Artificial Intelligence, Values, and Alignment'. *Minds and Machines* 30, no. 3 (1 September 2020): 411–37. https://doi.org/10.1007/s11023-020-09539-2.

[667] Aguirre, Anthony, Peter Bart Reiner, Harry Surden, and Gaia Dempsey. 'AI Loyalty by Design: A Framework for Governance of AI'. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 24 September 2021. https://papers.ssrn.com/abstract=3930338.; Benthall, Sebastian. 'For Safe AI Tomorrow, Fiduciary Duties for Big Tech Today'. dli-cornell-tech, 11 May 2022. https://www.dli.tech.cornell.edu/post/for-safe-ai-tomorrow-fiduciary-duties-for-big-tech-today.

[668] Nay, John. 'Law Informs Code: A Legal Informatics Approach to Aligning Artificial Intelligence with Humans'. SSRN Scholarly Paper. Rochester, NY, 13 September 2022. https://doi.org/10.2139/ssrn.4218031.

[669] Perry, Brandon, and Risto Uuk. 'AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk'. *Big Data and Cognitive Computing* 3, no. 2 (June 2019): 26. https://doi.org/10.3390/bdcc3020026.

[670] Leung, Jade. 'How Can We See the Impact of AI Strategy Research?' Presented at the EA Global: San Francisco 2019, 2019. https://forum.effectivealtruism.org/posts/Ae98k9d2gas32Yvmi/jade-leung-how-can-we-see-the-impact-of-ai-strategy-research.

[671] Critch, Andrew. 'Some AI Research Areas and Their Relevance to Existential Safety'. LessWrong, 19 November 2020. https://www.lesswrong.com/posts/hvGoYXi2kgnS3vxqb/some-ai-research-areas-and-their-relevance-to-existential-1.

→ Shape, provide, or spread narratives,[672] ideas, "memes,"[673] framings, or (legal) analogies[674] for AI governance.

→ Clarify or emphasize established principles within national law (e.g., precautionary principle and cost-benefit analysis[675]) and/or state obligations under international law (e.g., customary international law,[676] IHRL,[677] etc.).

3.3. Pathways to shaping court decisions

Shaping court decisions around AI systems that set critical precedent for the application of AI policy to advanced AI:

→ Advance legal scholarship with new arguments, interpretations, or analogies and metaphors for AI technology;[678]

→ Clarifying the "ordinary meaning" of key legal terms around AI;[679]

---

[672] Schiff, Daniel S. 'Setting the Agenda for AI: Actors, Issues, and Influence in United States Artificial Intelligence Policy'. Georgia Institute of Technology, 2022. https://osf.io/kw8xd/. (exploring the US policy context and arguing that "policy entrepreneurs can use persuasive narratives to influence legislators about AI policy, and that these narratives are just as effective as technical information. [...] [D]espite pervasive calls for public participation in AI governance, the public does not appear to play a key role in directing attention to AI's social and ethical implications nor in shaping concrete policy solutions, such that the emerging AI agenda remains primarily expert-driven.").

[673] Leung, Jade. 'How Can We See the Impact of AI Strategy Research?' Presented at the EA Global: San Francisco 2019, 2019. https://forum.effectivealtruism.org/posts/Ae98k9d2gas32Yvmi/jade-leung-how-can-we-see-the-impact-of-ai-strategy-research.

[674] See also Maas, Matthijs, 'AI is like… A literature review of AI metaphors and why they matter for policy.' *Institute for Law & AI*. AI Foundations Report 2. (October 2023). https://www.legalpriorities.org/research/ai-policy-metaphors; and see previously: Maas, 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'., pg. 215-216. For general work on how legal analogies can influence rulings made, see: Crootof, Rebecca. 'Autonomous Weapon Systems and the Limits of Analogy'. *Harvard National Security Journal* 9 (2018): 51–83. https://doi.org/10.2139/ssrn.2820727.; Lakier, Genevieve. 'The Problem Isn't the Use of Analogies but the Analogies Courts Use'. *Knight First Amendment Institute at Columbia University* (blog), 26 February 2018. https://knightcolumbia.org/content/problem-isnt-use-analogies-analogies-courts-use.

[675] Wiblin, Robert, and Keiran Harris. 'Carl Shulman on the Common-Sense Case for Existential Risk Work and Its Practical Implications'. 80,000 Hours Podcast. Accessed 11 October 2021. https://80000hours.org/podcast/episodes/carl-shulman-common-sense-case-existential-risks/.

[676] For an overview, see Rayfuse, Rosemary. 'Public International Law and the Regulation of Emerging Technologies'. In *The Oxford Handbook of Law, Regulation and Technology*, 2017. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-22. Pg. 503:
("the basic norms of international peace and security law, such as the prohibitions on the use of force and intervention in the domestic affairs of other states [...]; the basic principles of international humanitarian law, such as the requirements of humanity, distinction and proportionality [...]; the basic principles of international human rights law, including the principles of human dignity and the right to life, liberty, and security of the person [...]; and the basic principles of international environmental law, including the no-harm principle, the obligation to prevent pollution, the obligation to protect vulnerable ecosystems and species, the precautionary principle, and a range of procedural obligations relating to cooperation, consultation, notification, and exchange of information, environmental impact assessment, and participation [...]. The general customary rules on state responsibility and liability for harm also apply.").

[677] Vöneky, Silja. 'How Should We Regulate AI? Current Rules and Principles as Basis for "Responsible Artificial Intelligence"', 19 May 2020. https://papers.ssrn.com/abstract=3605440.

[678] Maas, Matthijs, 'AI is Like… A Literature Review of AI Metaphors and Why They Matter for Policy.' *Institute for Law & AI*. AI Foundations Report 2. (October 2023). https://www.legalpriorities.org/research/ai-policy-metaphors

[679] Martínez, Eric, and Christoph Winter. 'Ordinary Meaning of Existential Risk'. *LPP Working Paper No. 7-2022*, 2022. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4304670.

→ Judge seminars and training courses;[680]

→ Online information repositories.[681]

3.4. Pathways to shaping AI developers' decisions

Shaping individual lab decisions around AI governance:

→ Governmental regulations (e.g., industry risk, liability, criminal, etc.);

→ Institutional design choices: establish rules in the charter that enable the board of directors to make more cautious or pro-social choices,[682] and establish an internal AI ethics board[683] or internal audit functions;[684]

→ Campaigns or resources to educate researchers about AI risk, making AI safety research more concrete and legible, and/or creating common knowledge about researchers' perceptions of and attitudes towards these risks;[685]

→ Employee activism and pressure,[686] and documented communications of risks by employees (which make companies more risk averse because they are more likely to be held liable in court);[687]

→ Human rights norms generally applicable to business activities under the Ruggie Principles,[688] which amongst others can directly influence decisions by tech company oversight bodies;[689]

---

[680] Ash, Elliott, Daniel L. Chen, and Suresh Naidu. 'Ideas Have Consequences: The Impact of Law and Economics on American Justice'. Working Paper. Working Paper Series. National Bureau of Economic Research, February 2022. https://doi.org/10.3386/w29788. Discussed in: Matthews, Dylan, and Byrd Pinkerton. 'How a Resort Weekend for Judges Made Courts More Conservative'. Vox, 1 June 2019. https://www.vox.com/future-perfect/2019/6/1/18629859/judge-resort-weekend-naidu-manne-seminar-ginsburg.

[681] See generally Thompson, Neil, Brian Flanagan, Edana Richardson, Brian McKenzie, and Xueyun Luo. 'Trial by Internet: A Randomized Field Experiment on Wikipedia's Influence on Judges' Legal Reasoning'. SSRN Scholarly Paper. Rochester, NY, 27 July 2022. https://doi.org/10.2139/ssrn.4174200.

[682] Cihon, Peter, Jonas Schuett, and Seth D. Baum. 'Corporate Governance of Artificial Intelligence in the Public Interest'. *Information* 12, no. 7 (July 2021): 275. https://doi.org/10.3390/info12070275.

[683] Schuett, Jonas, Anka Reuel, and Alexis Carlier. 'How to Design an AI Ethics Board'. arXiv, 14 April 2023. https://doi.org/10.48550/arXiv.2304.07249.

[684] Schuett, Jonas. 'AGI Labs Need an Internal Audit Function'. arXiv, 26 May 2023. https://doi.org/10.48550/arXiv.2305.17038.

[685] Wasil, Akash, and Thomas Larsen. 'Ways to Buy Time'. LessWrong, 12 November 2022. https://www.lesswrong.com/posts/bkpZHXMJx3dG5waA7/ways-to-buy-time. I thank Zach Stein-Perlman for this suggestion.

[686] Belfield, Haydn. 'Activism by the AI Community: Analysing Recent Achievements and Future Prospects'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 15–21. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375814.

[687] Casper, Stephan. 'The 6D Effect: When Companies Take Risks, One Email Can Be Very Powerful.' EA Forum, 4 November 2023. https://forum.effectivealtruism.org/posts/QsfGEhFpMvgWjyusm/the-6d-effect-when-companies-take-risks-one-email-can-be. (referring to a "6D effect" of "the Duty to Due Diligence from Discoverable Documentation of Dangers").

[688] Vöneky, Silja. 'How Should We Regulate AI? Current Rules and Principles as Basis for "Responsible Artificial Intelligence"', 19 May 2020. https://papers.ssrn.com/abstract=3605440.

[689] Helfer, Laurence R., and Molly K. Land. 'The Facebook Oversight Board's Human Rights Future'. SSRN Scholarly Paper. Rochester, NY, 22 August 2022. https://doi.org/10.2139/ssrn.4197107.; Kulick, Andreas. 'Corporations as Interpreters and Adjudicators of International Human Rights Norms – Meta's Oversight Board and Beyond'. SSRN Scholarly Paper. Rochester, NY, 22 September 2022. https://papers.ssrn.com/abstract=4226521.; Wong, David, and Luciano Floridi. 'Meta's Oversight Board: A Review and Critical Assessment'. SSRN Scholarly Paper. Rochester, NY, 22 October 2022. https://papers.ssrn.com/abstract=4255817.

→ Develop and provide clear industry standards and resources for their implementation, such as AI risk management frameworks.[690]

Shaping industry-wide decisions around AI governance:

→ Governmental regulations (as above);

→ Ensure competition law frameworks enable cooperation on safety.[691]

3.5. Pathways to shaping AI research community decisions

Shaping AI research community decisions around AI governance:

→ Develop and disseminate clear guidelines and toolsets to facilitate responsible practices, such as:
  → Frameworks for pre-publication impact assessment of AI research;[692]
  → "Model cards" for the transparent reporting of benchmarked evaluations of a model's performance across conditions and for different groups;[693]
  → General risk management frameworks for evaluating and anticipating AI risks.[694]

→ Framing and stigmatization around decisions or practices;[695]

→ Participatory technology assessment processes.[696]

Shaping civil society decisions around AI governance:

→ Work with "gatekeeper" organizations to put issues on the advocacy agenda.[697]

---

[690] See for instance: ISO/IEC JTC 1/SC 42. 'ISO/IEC 23894:2023: Information Technology: Artificial Intelligence: Guidance on Risk Management'. ISO, February 2023. https://www.iso.org/standard/77304.html.; NIST. 'AI Risk Management Framework: AI RMF (1.0)'. Gaithersburg, MD: National Institute of Standards and Technology, 2023. https://doi.org/10.6028/NIST.AI.100-1; Barrett, Anthony, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, and Krystal Jackson. 'AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models'. Center for Long-Term Cybersecurity, November 2023. https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile/. And previously See also: Barrett, Anthony M., Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. 'Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks'. arXiv, 17 June 2022. https://doi.org/10.48550/arXiv.2206.08966.

[691] Hua, Shin-Shin, and Haydn Belfield. 'AI & Antitrust: Reconciling Tensions Between Competition Law and Cooperative AI Development'. *Yale Journal of Law and Technology* 23 (Spring 2021): 127. https://yjolt.org/ai-antitrust-reconciling-tensions-between-competition-law-and-cooperative-ai-development

[692] Ashurst, Carolyn, Markus Anderljung, Carina Prunkl, Jan Leike, Yarin Gal, Toby Shevlane, and Allan Dafoe. 'A Guide to Writing the NeurIPS Impact Statement'. *Centre for the Governance of AI (Medium)*, 19 May 2020. https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832.

[693] Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 'Model Cards for Model Reporting'. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–29, 2019. https://doi.org/10.1145/3287560.3287596.

[694] Barrett, Anthony, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, and Krystal Jackson. 'AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models'. Center for Long-Term Cybersecurity, November 2023. https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile/.

[695] Baum, Seth D. 'On the Promotion of Safe and Socially Beneficial Artificial Intelligence'. *AI & SOCIETY*, 28 September 2016. https://doi.org/10.1007/s00146-016-0677-0.; on some such forms of social pushback, see also informally lc. 'What an Actually Pessimistic Containment Strategy Looks Like'. LessWrong, 5 April 2022. https://www.lesswrong.com/posts/kipMvuaK3NALvFHc9/what-an-actually-pessimistic-containment-strategy-looks-like.

[696] Cremer, Carla Zoe, and Jess Whittlestone. 'Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI'. *International Journal of Interactive Multimedia and Artificial Intelligence* 6, no. 5 (2021): 100–109. https://www.ijimai.org/journal/sites/default/files/2021-02/ijimai_6_5_10.pdf

[697] Rosert, Elvira, and Frank Sauer. 'How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies'. *Contemporary Security Policy* 0, no. 0 (30 May 2020): 1–26. https://doi.org/10.1080/13523260.2020.1771508.

3.6. Pathways to shaping international institutions' decisions

Shaping international institutional decisions around AI governance:

- → Clarify global administrative law obligations;[698]
- → Influence domestic policy processes in order to indirectly shape transnational legal processes;[699]
- → Scientific expert bodies' role in informing multilateral treaty-making by preparing evidence for treaty-making bodies, scientifically advising these bodies, and directly exchanging with them at intergovernmental body sessions or dialogical events.[700]

Shaping standards bodies' decisions around AI governance:

- → Technical experts' direct participation in standards development;[701]
- → Advancing standardization of advanced AI-relevant safety best practices.[702]

3.7. Other pathways to shape various actors' decisions

Shaping various actors' decisions around AI governance:

- → Work to shape broad narratives around advanced AI, such as through compelling narratives or depictions of good outcomes;[703]
- → Work to shape analogies or metaphors used by the public, policymakers, or courts in thinking about (advanced) AI;[704]

[698] Benvenisti, Eyal. 'Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?' *European Journal of International Law* 29, no. 1 (23 July 2018): 9–82. https://doi.org/10.1093/ejil/chy031.

[699] See generally: Koh, Harold Hongju. 'Why Do Nations Obey International Law?' Edited by Abram Chayes, Antonia Handler Chayes, and Thomas M. Franck. *The Yale Law Journal* 106, no. 8 (1997): 2599–2659. https://doi.org/10.2307/797228.

[700] Orangias, Joseph. 'The Nexus between International Law and Science: An Analysis of Scientific Expert Bodies in Multilateral Treaty-Making'. *International Community Law Review* 25, no. 1 (1 April 2022): 60–93. https://doi.org/10.1163/18719732-bja10068.

[701] Ingersleben-Seip, Nora von. 'Competition and Cooperation in Artificial Intelligence Standard Setting: Explaining Emergent Patterns'. *Review of Policy Research* n/a, no. n/a. Accessed 25 January 2023. https://doi.org/10.1111/ropr.12538.

[702] Center for Long-Term Cybersecurity. 'Seeking Input and Feedback: AI Risk Management-Standards Profile for Increasingly Multi-Purpose or General-Purpose AI'. *CLTC* (blog), 2022. https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/; O'Keefe, Cullen, Jade Leung, and Markus Anderljung. 'How Technical Safety Standards Could Promote TAI Safety'. Effective Altruism Forum, 8 August 2022. https://forum.effectivealtruism.org/posts/zvbGXCxc5jBowCuNX/how-technical-safety-standards-could-promote-tai-safety.

[703] AI Impacts. 'AI Vignettes Project'. AI Impacts, 12 October 2021. https://aiimpacts.org/ai-vignettes-project/. Future of Life Institute. 'Content Sequence: Imagine A World'. *Future of Life Institute* (blog), 2023. https://futureoflife.org/content-sequence/imagine-a-world/.

[704] Maas, Matthijs, 'AI is Like… A Literature Review of AI Metaphors and Their Policy Effects.' *Institute for Law & AI*. AI Foundations Report #2. (October 2023). https://www.legalpriorities.org/research/ai-policy-metaphors

→ Pursue specific career paths with key actors to contribute to good policymaking.[705]

III. Prescriptive work: Identifying priorities and proposing policies

Finally, a third category of work aims to go beyond either analyzing the problem of AI governance (Part I) or surveying potential elements or options for governance solutions analytically (Part II). This category is rather prescriptive in that it aims to directly propose or advocate for specific policies or actions by key actors. This includes work focused on:

1. Articulating broad theories of change to identify priorities for AI governance (given a certain view of the problem and of the options available);
2. Articulating broad heuristics for crafting good AI regulation;
3. Putting forward policy proposals as well as assets that aim to help in their implementation.

1. Prioritization: Articulating theories of change

Achieving an understanding of the AI governance problem and potential options in response is valuable. Yet, this is not enough alone to deliver strategic clarity about which of these actors should be approached or which of these levers should be utilized in what ways. For that, it is necessary to develop more systematic accounts of different (currently held or possible) theories of change or impact.

The idea of exploring and comparing such theories of action is not new. There have been various accounts that aim to articulate the linkages between near-term actions and longer-term goals. Some of these have focused primarily on theories of change (or "impact") from the perspective of technical AI alignment.[706] Others have articulated more specific theories of impact for the advanced AI governance space.[707] These include:

---

[705] Brundage, Miles. 'Guide to Working in Artificial Intelligence Policy and Strategy'. 80,000 Hours, 13 June 2017. https://80000hours.org/articles/ai-policy-guide/.; For different country-level guides, see: Langosco, Lauro. 'AI Policy Careers in the EU'. EA Forum, 11 November 2019. https://forum.effectivealtruism.org/posts/XGPW25NZHq2WHbK9w/ai-policy-careers-in-the-eu.; Bowerman, Niel. 'The Case for Building Expertise to Work on US AI Policy'. 80,000 Hours, December 2020. https://80000hours.org/articles/us-ai-policy/.; 80,000 Hours. 'China-Related AI Safety and Governance Paths'. 80,000 Hours, February 2022. https://80000hours.org/career-reviews/china-related-ai-safety-and-governance-paths/.; Chua, Yi-Yang. 'Singapore AI Policy Career Guide'. EA Forum, 21 January 2021. https://forum.effectivealtruism.org/posts/umeMcbD4jDseLjsgT/singapore-ai-policy-career-guide. See also the careers reading guide in: BlueDot Impact. 'AI Governance Curriculum'. AI Safety Fundamentals, 2022. https://aisafetyfundamentals.com/ai-governance-curriculum (week 7).

[706] Dai, Wei. 'AI Safety "Success Stories"'. AI Alignment Forum, 7 September 2019. https://www.alignmentforum.org/posts/bnY3L48TtDrKTzGRb/ai-safety-success-stories.; Nanda, Neel. 'My Overview of the AI Alignment Landscape: A Bird's Eye View'. LessWrong, 16 December 2021. https://www.lesswrong.com/posts/SQ9cZtfrzDJmw9A2m/my-overview-of-the-ai-alignment-landscape-a-bird-s-eye-view.; Nanda, Neel. 'A Longlist of Theories of Impact for Interpretability'. LessWrong, 11 March 2022. https://www.lesswrong.com/posts/uK6sQCNMw8WKzJeCQ/a-longlist-of-theories-of-impact-for-interpretability. ; Hubinger, Evan. 'A Positive Case for How We Might Succeed at Prosaic AI Alignment'. LessWrong, 16 November 2021. https://www.lesswrong.com/posts/5ciYedyQDDqAcrDLr/a-positive-case-for-how-we-might-succeed-at-prosaic-ai.

[707] See also Aird, Michael, and Max Rauker. 'Survey on Intermediate Goals in AI Governance'. EA Forum, 17 March 2023. https://forum.effectivealtruism.org/posts/g4fXhiJyj6tdBhuBK/survey-on-intermediate-goals-in-ai-governance.

→ Dafoe's Asset-Decision model, which focuses on the direction of research activities to help (1) create assets which can eventually (2) inform impactful decisions;[708]

→ Leung's model for impactful AI strategy research that can shape key decisions by (1) those developing and deploying AI and (2) those actors shaping the environments in which it is developed and deployed (i.e., research lab environment, legislative environment, and market environment).[709]

→ Garfinkel's "AI Strategy: Pathways for Impact,"[710] which highlights three distinct pathways for positively influencing the development of advanced AI: (1) become a decision-maker (or close enough to influence one), (2) spread good memes that are picked up by decision-makers, and (3) think of good memes to spread and make them credible;

→ Baum's framework for "affecting the future of AI governance," which distinguishes several avenues by which AI policy could shape the long-term:[711] (1) improve current AI governance, (2) support AI governance communities, (3) advance research on future AI governance, (4) advance CS design of AI safety and ethics to create solutions, and (5) improve underlying governance conditions.

In addition, some have articulated specific scenarios for what successful policy action on advanced AI might look like,[712] especially in the relative near-term future ("AI strategy nearcasting").[713] However much further work is needed.

[708] Dafoe, Allan. 'AI Governance: Opportunity and Theory of Impact', 17 September 2020. https://www.allandafoe.com/opportunity. This model consists of a two-stage model for impact, which involves the direction of research activities to help (1) create assets ("technical solutions; strategic insights; shared perception of risks; a more cooperative worldview; well-motivated and competent advisors; credibility, authority, and connections for those experts"), which can eventually (2) inform impactful decisions ("by AI researchers, activists, public intellectuals, CEOs, generals, diplomats, or heads of state"). Notably, this model allows that there can be diverse views around which of the various assets or what breadth of assets are worth investing in today. Dafoe sketches a continuum between a narrow product model- and a broad field-building model of research and argues that while there is much current emphasis on delivering concrete research projects, given the uncertainty over advanced AI's technological trajectories and the prevailing political conditions around a future critical advanced AI moment, it is worth pursuing broad field-building activities for now. ("I believe the product model substantially underestimates the value of research in AI safety and, especially, AI governance; I estimate that the majority (perhaps ~80%) of the value of AI governance research comes from assets other than the narrow research product").

[709] Leung, Jade. 'How Can We See the Impact of AI Strategy Research?' 2019. https://forum.effectivealtruism.org/posts/Ae98k9d2gas32Yvmi/jade-leung-how-can-we-see-the-impact-of-ai-strategy-research. She argues that when approaching such decision-makers, one can aim to influence their (1) priorities, (2) strategies, and (3) tactics, and in doing so should (1) filter for making a case on a few tractable good things, (2) translate these into digestible memes, and (3) ensure your work reaches the key circle of influence.

[710] Garfinkel, Benjamin. 'AI Strategy: Pathways for Impact'. Accessed 6 April 2022.

[711] *Seth Baum on AI Governance*, 2021. https://www.youtube.com/watch?v=G-8uEg7mCdA.

[712] Hobbhahn, Marius, Max Räuker, Yannick Mühlhäuser, Jasper Götting, and Simon Grimm. 'What Success Looks Like'. Effective Altruism Forum, 28 June 2022. https://forum.effectivealtruism.org/posts/AuRBKFnjABa6c6GzC/what-success-looks-like. See for instance: Campos, Simon. 'AGI Timelines in Governance: Different Strategies for Different Timeframes'. EA Forum, 19 December 2022. https://forum.effectivealtruism.org/posts/Pt7MxstXxXHak4wkt/agi-timelines-in-governance-different-strategies-for.; Stein-Perlman, Zach. 'Framing AI Strategy'. AI Impacts, 6 February 2023. https://aiimpacts.org/ framing-ai-strategy/.

[713] See also Karnofsky, Holden. 'AI Strategy Nearcasting'. AI Alignment Forum, 25 August 2022. https://www.alignmentforum.org/posts/Qo2EkG3dEMv8GnX8d/ai-strategy-nearcasting. ("trying to answer key strategic questions about transformative AI, under the assumption that key events (e.g., the development of transformative AI) will happen in a world that is otherwise relatively similar to today's.").

2. General heuristics for crafting advanced AI policy

General heuristics for making policies relevant or actionable to advanced AI.

2.1. General heuristics for good regulation

Heuristics for crafting good AI regulation:

- → Utilizing and articulating suitable terminology for drafting and scoping AI regulations, especially risk-focused terms;[714]
- → Understand implications of different regulatory approaches (ex ante, ex post; risk regulation) for AI regulations;[715]
- → Grounding AI policy within an "all-hazards" approach to managing various other global catastrophic risks simultaneously;[716]
- → Requirements for an advanced AI regime to avoid "perpetual risk": exclusivity, benevolence, stability, and successful alignment;[717]
- → Establishing monitoring infrastructures to provide governments with actionable information.[718]

2.2. Heuristics for good institutional design

Heuristics for good institutional design:

- → General desiderata and tradeoffs for international institutional design in terms of questions of regime centralization or decentralization;[719]

---

[714] Maas, Matthijs, 'Concepts in advanced AI governance: A literature review of key terms and definitions.' *Institute for Law & AI*. AI Foundations Report 3. (October 2023). https://www.legalpriorities.org/research/advanced-ai-gov-concepts ; Schuett, Jonas. 'Defining the Scope of AI Regulations'. *Law, Innovation and Technology* 0, no. 0 (3 March 2023): 1–23. https://doi.org/10.1080/17579961.2023.2184135. See also Gutierrez, Carlos Ignacio, Anthony Aguirre, Risto Uuk, Claire C. Boine, and Matija Franklin. 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems'. Future of Life Institute - Working Paper, 5 October 2022. https://doi.org/10.2139/ssrn.4238951.

[715] Petit, Nicolas, and Jerome De Cooman. 'Models of Law and Regulation for AI'. EUI Working Paper RSCAS 2020/63. Social Science Research Network, 1 October 2020. https://doi.org/10.2139/ssrn.3706771.; Maas, Matthijs M. 'Aligning AI Regulation to Sociotechnical Change'. In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang. Oxford University Press, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.22. For a discussion of the implications, strengths, and shortcomings of a "risk regulation" approach to AI governance, see Kaminski, Margot E. 'Regulating the Risks of AI'. *Boston University Law Review* 103 (19 August 2022). https://doi.org/10.2139/ssrn.4195066.

[716] Sepasspour, Rumtin. 'All-Hazards Policy for Global Catastrophic Risk'. Technical Report. Global Catastrophic Risk Institute, 2 November 2023. https://gcrinstitute.org/all-hazards-policy/.

[717] Casper, Stephen. 'Avoiding Perpetual Risk from TAI'. LessWrong, 26 December 2022. https://www.lesswrong.com/posts/FfTxEf3uFPsZf9EMP/avoiding-perpetual-risk-from-tai.

[718] Ho, Anson. 'Future-Proof: Monitoring the Development, Deployment, and Impacts of Artificial Intelligence'. *Journal of Science Policy & Governance* 22, no. 03 (11 September 2023). http://www.sciencepolicyjournal.org/article_1038126_jspg220305.html.; Whittlestone, Jess, and Jack Clark. 'Why and How Governments Should Monitor AI Development'. *ArXiv:2108.12427 [Cs]*, 31 August 2021. http://arxiv.org/abs/2108.12427.

[719] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890.; Cihon, Peter, Matthijs M Maas, and Luke Kemp. 'Should Artificial Intelligence Governance Be Centralised? Six Design Lessons from History'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 11, 2019. https://www.cser.ac.uk/media/uploads/files/Cihon_et_al-_2019-_Should_AI_Governance_be_Centralised.pdf.

→ Procedural heuristics for organizing international negotiation processes: ensure international AI governance fora are inclusive of Global South actors;[720]

→ Ideal characteristics of global governance systems for high-risk AI, such as those that (1) govern dual-use technology; (2) take a risk-based approach; (3) provide safety measures; (4) incorporate technically informed, expert-driven, multi-stakeholder processes that enable rapid iteration; (5) where the effects are consistent with the treaty's intent; and (6) that possess enforcement mechanisms.[721]

2.3. Heuristics for future-proofing governance

Heuristics for future-proofing governance regimes and desiderata and systems for making existing regulations more adaptive, scalable, or resilient:[722]

→ Traditional (treaty) reform or implementation mechanisms:
  → The formal treaty amendment process;[723]
  → Unilateral state actions (explanatory memoranda and treaty reservations) or multilateral responses (Working Party Resolution) to adapt multilateral treaties;[724]
  → The development of lex scripta treaties through the lex posteriori of customary international law, spurred by new state behavior.[725]

→ Adaptive treaty interpretation methods:
  → Evolutionary interpretation of treaties;[726]

[720] Adan, Sumaya Nur. 'The Case for Including the Global South in AI Governance Discussions'. GovAI Blog, 20 October 2023. https://www.governance.ai/post/the-case-for-including-the-global-south-in-ai-governance-conversations. Abungu, Cecil, Michelle Malonza, and Sumaya Nur Adan. 'Can Apparent Bystanders Distinctively Shape An Outcome? The Extent To Which Some Global South Countries Could Matter in the Global Catastrophic Risk-Focused Governance of Artificial Intelligence Development'. ILINA STAI Paper, 2023 forthcoming.

[721] See the framework set out in: Llerena, Stephan. 'Global Governance of High-Risk Artificial Intelligence', 27 October 2023. (draft manuscript).

[722] See generally: Stauffer, Maxime, Malou Estier, Konrad Seifert, and Jacob Arbeid. 'The FAIR Framework - A Future-Proofing Methodology'. Simon Institute for Longterm Governance, 26 April 2023. https://www.simoninstitute.ch/blog/post/the-fair-framework-a-future-proofing-methodology/.; and previously Chander, Anupam. 'Future-Proofing Law'. *UC Davis Law Review* 1, no. 51 (2017). https://lawreview.law.ucdavis.edu/issues/51/1/Symposium/51-1_Chander.pdf.; Ranchordás, Sofia, and Mattis van't Schip. 'Future-Proofing Legislation for the Digital Age'. In *Time, Law, and Change : An Interdisciplinary Study*, edited by Sofia Ranchordás and Yaniv Roznai, 1st ed., 347–66. Oxford: Hart Publishing, 2020. http://www.bloomsburycollections.com/book/time-law-and-change-an-interdisciplinary-study/ch16-future-proofing-legislation-for-the-digital-age/.

[723] Bowman, M. J. 'The Multilateral Treaty Amendment Process—A Case Study'. *International & Comparative Law Quarterly* 44, no. 3 (July 1995): 540–59. https://doi.org/10.1093/iclqaj/44.3.540.

[724] Smith, Bryant Walker. 'New Technologies and Old Treaties'. *AJIL Unbound* 114 (ed 2020): 152–57. https://doi.org/10.1017/aju.2020.28.

[725] Crootof, Rebecca. 'Change Without Consent: How Customary International Law Modifies Treaties'. *Yale Journal of International Law* 41, no. 2 (2016): 65. https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1670&context=yjil

[726] See generally: Vidigal, Geraldo. 'Evolutionary Interpretation and International Law'. *Journal of International Economic Law* 24, no. 1 (1 March 2021): 203–19. https://doi.org/10.1093/jiel/jgaa035.; Abi-Saab, Georges, Kenneth Keith, Gabrielle Marceau, and Clément Marquet, eds. *Evolutionary Interpretation and International Law*. S.l.: Hart Publishing, 2021.

- → Treaty interpretation under the principle of systemic integration.[727]

→ Instrument choices that promote flexibility:
- → Use of framework conventions;[728]
- → Use of informal governance institutions;[729]
- → The subsequent layering of soft law on earlier hard-law regimes;[730]
- → Use of uncorrelated governance instruments to enable legal resilience.[731]

→ Regime design choices that promote flexibility:
- → Scope: include key systems ("general-purpose AI systems," "highly capable foundation models," "frontier AI systems," etc.) within the material scope of the regulation;[732]
- → Phrasing: in-text technological neutrality or deliberate ambiguity;[733]
- → Flexibility provisions: textual flexibility provisions[734] such as exceptions or flexibility clauses.

→ Flexibility approaches beyond the legal regime:
- → Pragmatic and informal "emergent flexibility" about the meaning of norms and rules during crises.[735]

---

[727] See generally: Mclachlan, Campbell. 'The Principle of Systemic Integration and Article 31(3)(c) of the Vienna Convention'. *International and Comparative Law Quarterly* 54, no. 2 (April 2005): 279–320. https://doi.org/10.1093/iclq/lei001.; Aspremont, Jean d'. 'The Systemic Integration of International Law by Domestic Courts: Domestic Judges as Architects of the Consistency of the International Legal Order'. In *The Practice of International and National Courts and the (De-)Fragmentation of International Law*, edited by A. Nollkaemper and O.K. Fauchald. Hart, 2012. https://papers.ssrn.com/abstract=1401019. Van Aaken. 'Defragmentation of Public International Law Through Interpretation: A Methodological Proposal'. *Indiana Journal of Global Legal Studies* 16, no. 2 (2009): 483. https://doi.org/10.2979/gls.2009.16.2.483.; Peters, Anne. 'The Refinement of International Law: From Fragmentation to Regime Interaction and Politicization'. *International Journal of Constitutional Law* 15, no. 3 (30 October 2017): 671–704. https://doi.org/10.1093/icon/mox056.

[728] Matz-Lück, Nele. 'Framework Conventions as a Regulatory Tool'. *Goettingen Journal of International Law* 3 (2009): 439–58. https://doi.org/10.3249/1868-1581-1-3-MATZ-LUECK.

[729] Morin, Jean‐Frédéric, Hugo Dobson, Claire Peacock, Miriam Prys‐Hansen, Abdoulaye Anne, Louis Bélanger, Peter Dietsch, et al. 'How Informality Can Address Emerging Issues: Making the Most of the G7'. *Global Policy* 10, no. 2 (May 2019): 267–73. https://doi.org/10.1111/1758-5899.12668.

[730] Israel, Brian. 'Treaty Stasis'. *AJIL Unbound* 108 (ed 2014): 63–69. https://doi.org/10.1017/S2398772300001860.

[731] Marchant, Gary E, and Yvonne A Stevens. 'Resilience: A New Tool in the Risk Governance Toolbox for Emerging Technologies'. *U.C. Davis Law Review* 51, no. 1 (2017): 233–71. https://lawreview.law.ucdavis.edu/issues/51/1/Symposium/51-1_Marchant_Stevens.pdf

[732] Uuk, Risto. 'General Purpose AI and the AI Act'. Future of Life Institute, May 2022. https://artificialintelligenceact.eu/wp-content/uploads/2022/05/General-Purpose-AI-and-the-AI-Act.pdf. See also: Gutierrez, Carlos Ignacio, Anthony Aguirre, Risto Uuk, Claire C. Boine, and Matija Franklin. 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems'. Future of Life Institute - Working Paper, 5 October 2022. https://doi.org/10.2139/ssrn.4238951.

[733] See Canfil, Justin Key. 'Yesterday's Reach: How International Law Keeps Pace with Technological Change'. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 2 January 2020. https://papers.ssrn.com/abstract=3684991.

[734] Koremenos, Barbara. *The Continent of International Law: Explaining Agreement Design*. Cambridge: Cambridge University Press, 2016. https://doi.org/10.1017/CBO9781316415832.; Helfer, Laurence R. 'Flexibility in International Agreements'. In *Interdisciplinary Perspectives on International Law and International Relations*, edited by Jeffrey L. Dunoff and Mark A. Pollack, 175–96. Cambridge: Cambridge University Press, 2012. https://doi.org/10.1017/CBO9781139107310.010.; Boockmann, B., and Paul W. Thurner. 'Flexibility Provisions in Multilateral Environmental Treaties'. *International Environmental Agreements: Politics, Law and Economics* 6, no. 2 (1 June 2006): 113–35. https://doi.org/10.1007/s10784-006-9001-7.

[735] Búzás, Zoltán I, and Erin R Graham. 'Emergent Flexibility in Institutional Development: How International Rules Really Change'. *International Studies Quarterly* 64, no. 4 (7 December 2020): 821–33. https://doi.org/10.1093/isq/sqaa049.

3. Policy proposals, assets and products

That is, what are specific proposals for policies to be implemented? How can these proposals serve as products or assets in persuading key actors to act upon them?

In this context, a "(decision-relevant) asset" can be defined as: "resources that can be used by other actors in pursuing *pathways* to influence *key actors* with the aim to induce how these key actors make *key decisions* (e.g., about whether or how to use their *levers*). This includes new technical research insights, worked-out policy *products*, networks of direct advocacy, memes, or narratives."

A "(policy) product" can be defined as "a subclass of *assets*; specific legible proposals that can be presented to *key actors*."

Specific proposals for advanced AI-relevant policies; note that these are presented without comparison or prioritization. This list is non-exhaustive. Many proposals moreover combine several ideas, falling into different categories.

3.1. Overviews and collections of policies

→ Previous collections of older proposals, such as Dewey's list of "long-term strategies for ending existential risk"[736] as well as Sotala and Yampolskiy's survey of high-level "responses" to AI risk.[737]

→ More recent lists and collections of proposed policies to improve the governance, security, and safety of AI development[738] in domains such as compute security and governance; software export controls; licenses;[739] policies to establish improved standards, system evaluations, and licensing regimes; procurement rules and funding for AI safety;[740] or to establish a multinational AGI consortium to enable oversight of advanced AI, a global compute cap, and affirmative safety evaluations.[741]

---

[736] Dewey, Daniel. 'Long-Term Strategies for Ending Existential Risk from Fast Takeoff'. In *Risks of Artificial Intelligence*. Chapman and Hall/CRC, 2015. https://www.taylorfrancis.com/chapters/edit/10.1201/b19187-14/long-term-strategies-ending-existential-risk-fast-takeoff-daniel-dewey. (including international coordination, sovereign AI, AI-empowered project, and decisive technological advantage).

[737] Sotala, Kaj, and Roman Yampolskiy. 'Responses to the Journey to the Singularity'. In *The Technological Singularity*, edited by Victor Callaghan, James Miller, Roman Yampolskiy, and Stuart Armstrong, 25–83. The Frontiers Collection. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017. https://doi.org/10.1007/978-3-662-54033-6_3.; and also Sotala, Kaj, and Roman V. Yampolskiy. 'Responses to Catastrophic AGI Risk: A Survey.' Technical Report. Berkeley, CA: Machine Intelligence Research Institute, 2013. https://intelligence.org/files/ResponsesAGIRisk.pdf. (in particular, "societal proposals," including: "do nothing," "integrate with society," "regulate research," "enhance human capabilities," and "relinquish technology").

[738] See generally Stein-Perlman, Zach. 'List of Lists of Government AI Policy Ideas'. EA Forum, 17 April 2023. https://forum.effectivealtruism.org/posts/wkAoqnaP7DhqHjyzh/list-of-lists-of-government-ai-policy-ideas.

[739] Muelhauser, Luke. '12 Tentative Ideas for US AI Policy'. *Open Philanthropy* (blog), 17 April 2023. https://www.openphilanthropy.org/research/12-tentative-ideas-for-us-ai-policy/.

[740] Hashim, Shakeel. 'Proposals for AI Regulation'. AI Safety Communications Centre, 7 September 2023. https://aiscc.org/2023/09/07/proposals-for-ai-regulation/.

[741] Miotti, Andrea, and Akash Wasil. 'Taking Control: Policies to Address Extinction Risks from Advanced AI'. arXiv, 31 October 2023. https://doi.org/10.48550/arXiv.2310.20563. For another comparison, see also Future of Life Institute. 'AI Governance Scorecard and Safety Standards Policy: Evaluating Proposals for AI Governance and Providing a Regulatory Framework for Robust Safety Standards, Measures and Oversight'. Future of Life Institute, October 2023. https://futureoflife.org/project/uk-ai-safety-summit/.

3.2. Proposals to regulate AI using existing authorities, laws, or institutions

In particular, drawing on evaluations of the default landscape of regulations applied to AI (see Section I.3.3), and of the levers of governance for particular governments (see Section II.2.4).

Regulate AI using existing laws or policies

- → Strengthen or reformulate existing laws and policies, such as EU competition law,[742] contract and tort law,[743] etc.;
- → Strengthen or reorganize existing international institutions[744] rather than establishing new institutions;[745]
- → Extend or apply existing principles and regimes in international law,[746] including, amongst others:
  - → Norms of international peace and security law:
    - → Prohibitions on the use of force and intervention in the domestic affairs of other states;
    - → Existing export control and nonproliferation agreements.
  - → Principles of international humanitarian law, such as:
    - → Distinction and proportionality in wartime;
    - → Prohibition on weapons that are by nature indiscriminate or cause unnecessary suffering;
    - → The requirements of humanity;
    - → The obligation to conduct legal reviews of new weapons or means of war (Article 36 under Additional Protocol I to the Geneva Conventions).
  - → Norms of international human rights law[747] and human rights and freedoms, including the right to life and freedom from cruel, inhuman, and degrading treatment, among others; the rights to freedom of expression, association, and security of the person, among others; and the principle of human dignity;[748]

---

[742] Hua, Shin-Shin, and Haydn Belfield. 'Effective Enforceability of EU Competition Law Under Different AI Development Scenarios: A Framework for Legal Analysis'. *Verfassungsblog* (blog), 18 August 2022. https://verfassungsblog.de/effective-enforceability-of-eu-competition-law-under-different-ai-development-scenarios/.

[743] Boine, Claire. "Artificial intelligence and civil liability in the European Union." In Artificial Intelligence Law: between sectorial and general rules. Comparative perspectives. 6/2023. Bruylant.

[744] Roberts, Huw, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. 'Global AI Governance: Barriers and Pathways Forward'. SSRN Scholarly Paper. Rochester, NY, 29 September 2023. https://doi.org/10.2139/ssrn.4588040. Sepasspour, Rumtin. 'A Reality Check and a Way Forward for the Global Governance of Artificial Intelligence'. *Bulletin of the Atomic Scientists*, 10 September 2023. https://www.tandfonline.com/doi/abs/10.1080/00963402.2023.2245249.

[745] Roberts, Huw. 'Opinion – A New International AI Body Is No Panacea'. *E-International Relations* (blog), 11 August 2023. https://www.e-ir.info/2023/08/11/opinion-a-new-international-ai-body-is-no-panacea/.

[746] Kunz, Martina, and Seán Ó hÉigeartaigh. 'Artificial Intelligence and Robotization'. In *Oxford Handbook on the International Law of Global Security*, edited by Robin Geiss and Nils Melzer. Oxford University Press, 2021. https://papers.ssrn.com/abstract=3310421.; Vöneky, Silja. 'How Should We Regulate AI? Current Rules and Principles as Basis for "Responsible Artificial Intelligence"', 19 May 2020. https://papers.ssrn.com/abstract=3605440. For a broader review of international norms applicable to emerging existential risks, including those from technology, see: Villalobos, José Jaime, Matthijs Maas, and Christoph Winter. 'States Must Mitigate Existential Risk under International Law', Institute for Law & AI working paper. (Forthcoming).

[747] Vöneky, Silja. 'How Should We Regulate AI? Current Rules and Principles as Basis for "Responsible Artificial Intelligence"', 19 May 2020. https://papers.ssrn.com/abstract=3605440.; For an interesting angle: Bajgar, Ondrej, and Jan Horenovsky. 'Negative Human Rights as a Basis for Long-Term AI Safety and Regulation'. *Journal of Artificial Intelligence Research*, 2022, 30. https://arxiv.org/abs/2208.14788

[748] Though this is only recognized by some courts.

→ Norms of international environmental law, including the no-harm principle and the principle of prevention and precaution;

→ International criminal law, with regard to war crimes and crimes against humanity and with regard to case law of international criminal courts regarding questions of effective control;[749]

→ Rules on state responsibility,[750] including state liability for harm;

→ Peremptory norms of jus cogens, outlawing, for example, genocide, maritime piracy, slavery, wars of aggression, and torture;

→ International economic law:[751] security exception measures under international trade law and non-precludement measures under international investment law, amongst others;[752]

→ International disaster law: obligations regarding disaster preparedness, including forecasting and pre-disaster risk assessment, multi-sectoral forecasting and early warning systems, disaster risk and emergency communication mechanisms, etc. (Sendai Framework);

→ Legal protections for the rights of future generations: including existing national constitutional protections for the rights of future generations[753] and a potential future UN Declaration on Future Generations.[754]

Proposals to set soft-law policy through existing international processes

→ Proposals for engagement in existing international processes on AI: support the campaign to ban lethal autonomous weapons systems,[755] orchestrate soft-law policy under G20,[756] engage in debate about digital technology governance at the UN Summit for the Future,[757] etc.

3.3. Proposals for new policies, laws, or institutions

A range of proposals for novel policies.

[749] Burri, Thomas. 'International Law and Artificial Intelligence'. *German Yearbook of International Law* 60 (27 October 2017): 91–108. http://dx.doi.org/10.2139/ssrn.3060191

[750] See also: Boutin, Bérénice. 'State Responsibility in Relation to Military Applications of Artificial Intelligence'. *Leiden Journal of International Law* 36, no. 1 (March 2023): 133–50. https://doi.org/10.1017/S0922156522000607.

[751] Liu, Han-Wei, and Ching-Fu Lin. 'Artificial Intelligence and Global Trade Governance: A Pluralist Agenda'. *Harvard International Law Journal* 61, no. 2 (2020). https://papers.ssrn.com/abstract=3675505.

[752] See generally: McLaughlin, Mark. 'Regulating Artificial Intelligence in International Investment Law'. *The Journal of World Investment & Trade* 24, no. 2 (5 April 2023): 256–300. https://doi.org/10.1163/22119000-12340288.

[753] Araújo, Renan, and Leonie Koessler. 'The Rise of the Constitutional Protection of Future Generations'. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 30 September 2021. https://papers.ssrn.com/abstract=3933683.

[754] Hale, Thomas, Finlay Moorhouse, Toby Ord, and Anne-Marie Slaughter. 'Toward a Declaration on Future Generations', 12 January 2023. https://www.bsg.ox.ac.uk/research/publications/toward-declaration-future-generations.

[755] Aguirre, Anthony. 'Why Those Who Care about Catastrophic and Existential Risk Should Care about Autonomous Weapons'. EA Forum, 11 November 2020. https://forum.effectivealtruism.org/posts/oR9tLNRSAep293rr5/why-those-who-care-about-catastrophic-and-existential-risk-2.

[756] Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. 'Policy Brief: The Creation of a G20 Coordinating Committee for the Governance of Artificial Intelligence'. *AI and Ethics*, 6 October 2020. https://doi.org/10.1007/s43681-020-00019-y.

[757] Moorhouse, Fin, and Avital Balwit. 'Major UN Report Discusses Existential Risk and Future Generations (Summary)'. EA Forum, 17 September 2021. https://forum.effectivealtruism.org/posts/Fwu2SLKeM5h5v95ww/major-un-report-discusses-existential-risk-and-future.

Impose (temporary) pauses on AI development

→ Coordinated pauses amongst AI developers whenever they identify hazardous capabilities;[758]

→ Temporary pause on large-scale system training beyond a key threshold,[759] giving time for near-term policy-setting in domains such as robust third-party auditing and certification, regulation of access to computational power, establishment of capable national AI agencies, and establishment of liability for AI-caused harms, etc.;[760]

→ (Permanent) moratoria on developing (certain forms of) advanced AI.[761]

Establish licensing regimes

→ Evaluation and licensing regimes: establishment of a AI regulation regime for frontier AI systems, comprising "(1) standard-setting processes to identify appropriate requirements for frontier AI developers, (2) registration and reporting requirements to provide regulators with visibility into frontier AI development processes, and (3) mechanisms to ensure compliance with safety standards for the development and deployment of frontier AI models."[762]

Establish lab-level safety practices

→ Proposals for establishing corporate governance and soft law: establish Responsible Scaling Policies (RSPs)[763] and establish corporate governance and AI certification schemes.[764]

Establish governance regimes on AI inputs (compute, data)

→ Compute governance regimes: establish on-chip firmware mechanisms, inspection regimes, and supply chain monitoring and custody mechanisms to ensure no actor can use large quantities of specialized chips to execute ML training runs in violation of established rules;[765]

---

[758] Alaga, Jide, and Jonas Schuett. 'Coordinated Pausing: An Evaluation-Based Coordination Scheme for Frontier AI Developers'. arXiv, 30 September 2023. https://doi.org/10.48550/arXiv.2310.00374.

[759] Future of Life Institute. 'Pause Giant AI Experiments: An Open Letter'. *Future of Life Institute* (blog), 30 March 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/. PauseAI. 'PauseAI Proposal'. Accessed 28 August 2023. https://pauseai.info/proposal.; See also Bilge, Tolga. 'Treaty on Artificial Intelligence Safety and Cooperation (TAISC)', 2023. https://taisc.org/taisc ("States Parties shall prohibit Large Training Runs under their jurisdiction or control and shall not assist, encourage or induce, in any way, anyone to engage in conducting Large Training Runs, with the exception of training runs conducted by the Joint AI Safety Laboratory.").

[760] Future of Life Institute. 'Policymaking in the Pause: What Can Policymakers Do Now to Combat Risks from Advanced AI Systems?' Future of Life Institute, 12 April 2023. https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf.

[761] Aguirre, Anthony. 'Close the Gates to an Inhuman Future: How and Why We Should Choose to Not Develop Superhuman General-Purpose Artificial Intelligence'. SSRN Scholarly Paper. Rochester, NY, 20 October 2023. https://papers.ssrn.com/abstract=4608505.

[762] Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, et al. 'Frontier AI Regulation: Managing Emerging Risks to Public Safety'. arXiv, 11 July 2023. https://doi.org/10.48550/arXiv.2307.03718.

[763] ARC Evals. 'Responsible Scaling Policies (RSPs)'. ARC Evals, 26 September 2023. https://evals.alignment.org/blog/2023-09-26-rsp/.; Anthropic. 'Anthropic's Responsible Scaling Policy, Version 1.0', 19 September 2023. http://anthropic.com/responsible-scaling-policy

[764] Cihon, Peter, Jonas Schuett, and Seth D. Baum. 'Corporate Governance of Artificial Intelligence in the Public Interest'. *Information* 12, no. 7 (July 2021): 275. https://doi.org/10.3390/info12070275.; Cihon, Peter, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum. 'AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries'. *IEEE Transactions on Technology and Society* 2, no. 4 (December 2021): 200–209. https://doi.org/10.1109/TTS.2021.3077595.

[765] Shavit, Yonadav. 'What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training Via Compute Monitoring', 2023. https://paperswithcode.com/paper/what-does-it-take-to-catch-a-chinchilla.

→ Data governance: establish public data trusts to assert control over public training data for foundation models.[766]

Establish domestic institutions for AI governance

→ Proposals for new domestic institutions: US "AI Control Council"[767] or National Algorithms Safety Board,[768] and European AI Agency[769] or European AI Office.[770]

Establish international AI research consortia

Proposals to establish new international hubs or organizations aimed at AI research.[771]

→ A diverse range of proposals for international institutions, including: a "CERN for AI,"[772] "European Artificial Intelligence megaproject,"[773] "Multilateral AI Research Institute (MAIRI),"[774] "international large-scale AI R&D projects,"[775] a collaborative UN superintelligence research project,[776] "international organization that could serve as clearing-house for research into AI,"[777] "joint international AI project with a monopoly on hazardous AI development,"[778] "UN AI Research

[766] Chan, Alan, Herbie Bradley, and Nitarshan Rajkumar. 'Reclaiming the Digital Commons: A Public Data Trust for Training Data'. arXiv, 15 March 2023. https://doi.org/10.48550/arXiv.2303.09001.

[767] Korinek, Anton. 'Why We Need a New Agency to Regulate Advanced Artificial Intelligence: Lessons on AI Control from the Facebook Files'. *Brookings* (blog), 8 December 2021. https://www.brookings.edu/research/why-we-need-a-new-agency-to-regulate-advanced-artificial-intelligence-lessons-on-ai-control-from-the-facebook-files/.

[768] Ben Shneiderman, opinion contributor. 'Do We Need a National Algorithms Safety Board?' Text. *The Hill* (blog), 28 February 2023. https://thehill.com/opinion/technology/3876569-do-we-need-a-national-algorithms-safety-board/.

[769] Stix, Charlotte. 'Foundations for the Future: Institution Building for the Purpose of Artificial Intelligence Governance'. *AI and Ethics*, 29 September 2021. https://doi.org/10.1007/s43681-021-00093-w.

[770] Curtis, Sam, Felicity Reddel, and Nicolas Moës. 'A Blueprint for the European AI Office'. The Future Society, 17 October 2023. https://thefuturesociety.org/a-blueprint-for-the-european-ai-office/.

[771] A more detailed review of some of these can be found in: Maas, Matthijs, and Villalobos, José Jaime. 'International AI institutions: A literature review of models, examples, and proposals.' *Institute for Law & AI*, AI Foundations Report 1. (September 2023). https://www.legalpriorities.org/research/international-ai-institutions

[772] Fischer, Sophie-Charlotte, and Andreas Wenger. 'A Politically Neutral Hub for Basic AI Research'. Policy Perspectives. Zurich: CSS, ETH Zurich, March 2019. http://www.css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP7-2_2019-E.pdf.

[773] Stix, Charlotte. 'An Infrastructural Framework to Achieve a European Artificial Intelligence Megaproject', 30 September 2019. https://www.researchgate.net/publication/340574784_An_infrastructural_framework_to_achieve_a_European_artificial_intelligence_megaproject.

[774] Zhang, Daniel, Christie Lawrence, Michael Sellitto, Russell Wald, Marietje Schaake, Daniel E. Ho, Russ Altman, and Andrew Grotto. 'Enhancing International Cooperation in AI Research: The Case for a Multilateral AI Research Institute'. Stanford Institute for Human-Centered Artificial Intelligence, May 2022. https://hai.stanford.edu/white-paper-enhancing-international-cooperation-ai-research-case-multilateral-ai-research-institute.

[775] Kerry, Cameron F, Joshua P Meltzer, and Andrea Renda. 'AI Cooperation on the Ground: AI Research and Development on a Global Scale'. Brookings Institute & Forum for Cooperation on Artificial Intelligence (FCAI), October 2022. https://www.brookings.edu/wp-content/uploads/2022/11/FCAI-October-2022.pdf.

[776] Castel, J.G., and Mathew E. Castel. 'The Road to Artificial Superintelligence - Has International Law a Role to Play?' *Canadian Journal of Law & Technology* 14 (2016). https://ojs.library.dal.ca/CJLT/article/download/7211/6256. (pg 11-12).

[777] Neufville, Robert de, and Seth D. Baum. 'Collective Action on Artificial Intelligence: A Primer and Review'. *Technology in Society* 66 (1 August 2021): 101649. https://doi.org/10.1016/j.techsoc.2021.101649.

[778] Dewey, Daniel. 'Long-Term Strategies for Ending Existential Risk from Fast Takeoff'. In *Risks of Artificial Intelligence*. Chapman and Hall/CRC, 2015. https://www.taylorfrancis.com/chapters/edit/10.1201/b19187-14/long-term-strategies-ending-existential-risk-fast-takeoff-daniel-dewey. Pg. 7.

Organization,"[779] a "good-faith joint US-China AGI project,"[780] "AI for shared prosperity,"[781] and a proposal for a new "Multinational AGI Consortium."[782]

Establish bilateral agreements and dialogues

→ Establish confidence-building measures[783] and pursue international AI safety dialogues.[784]

Establish multilateral international agreements

Proposal to establish a new multilateral treaty on AI:[785]

→ "Treaty on Artificial Intelligence Safety and Cooperation (TAISC),"[786] global compute cap treaty,[787] "AI development convention,"[788] "Emerging Technologies Treaty,"[789] "Benevolent AGI Treaty,"[790] "pre-deployment agreements,"[791] and many other proposals.

Establish international governance institutions

Proposals to establish a new international organization, along one or several models:[792]

→ A diverse range of proposals for international institutions, including a Commission on Frontier AI, an Advanced AI Governance Organization, a Frontier AI Collaborative, and an AI Safety Project;[793] an

[779] Kemp, Luke, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigeartaigh. 'UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance'. Centre for the Study of Existential Risk and Leverhulme Centre for the Future of Intelligence, 26 February 2019. https://www.cser.ac.uk/news/advice-un-high-level-panel-digital-cooperation/.

[780] Bensinger, Rob. 'Ngo's View on Alignment Difficulty'. Machine Intelligence Research Institute, 15 December 2021. https://intelligence.org/2021/12/14/ngos-view-on-alignment-difficulty/.

[781] Cave, Stephen, and Seán S. ÓhÉigeartaigh. 'An AI Race for Strategic Advantage: Rhetoric and Risks'. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 36–40. New Orleans LA USA: ACM, 2018. https://doi.org/10.1145/3278721.3278780.

[782] Hausenloy, Jason, Andrea Miotti, and Claire Dennis. 'Multinational AGI Consortium (MAGIC): A Proposal for International Coordination on AI'. arXiv, 13 October 2023. https://doi.org/10.48550/arXiv.2310.09217.

[783] Shoker, Sarah, Andrew Reddie, Sarah Barrington, Ruby Booth, Miles Brundage, Husanjot Chahal, Michael Depp, et al. 'Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings'. arXiv, 3 August 2023. https://doi.org/10.48550/arXiv.2308.00862.

[784] Guest, Oliver, Michael Aird, and Fynn Heide. 'International AI Safety Dialogues: Benefits, Risks, and Best Practices'. Institute for AI Policy and Strategy (IAPS), 31 October 2023. https://www.iaps.ai/research/international-ai-safety-dialogues.

[785] ibid. pg. 43-44 ('Multilateral AI treaties without institutions'). Note that in some (but not all) cases these treaty proposals envisage the establishment of a new international institution.

[786] Bilge, Tolga. 'Treaty on Artificial Intelligence Safety and Cooperation (TAISC)', 2023. https://taisc.org.

[787] Miotti, Andrea, and Akash Wasil. 'An International Treaty to Implement a Global Compute Cap for Advanced Artificial Intelligence'. SSRN Scholarly Paper. Rochester, NY, 30 October 2023. https://doi.org/10.2139/ssrn.4617094.

[788] Dewey, Daniel. 'Long-Term Strategies for Ending Existential Risk from Fast Takeoff'. In *Risks of Artificial Intelligence*. Chapman and Hall/CRC, 2015. https://www.taylorfrancis.com/chapters/edit/10.1201/b19187-14/long-term-strategies-ending-existential-risk-fast-takeoff-daniel-dewey. Pg. 7-8.

[789] Wilson, Grant. 'Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law'. *Va. Envtl. LJ* 31 (2013): 307. http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/velj31&section=12

[790] Ramamoorthy, Anand, and Roman Yampolskiy. 'Beyond MAD?: The Race for Artificial General Intelligence'. *ITU JOURNAL: ICT DISCOVERIES* 1, no. 1 (2018): 8. https://www.itu.int/en/journal/001/Documents/itu2018-9.pdf

[791] Belfield, Haydn. 'The Rival AI Deployment Problem: A Pre-Deployment Agreement as the Least-Bad Response'. EA Forum, 23 September 2022. https://forum.effectivealtruism.org/posts/uSH6DqjzggAYQGjxm/the-rival-ai-deployment-problem-a-pre-deployment-agreement.

[792] A more detailed review can be found in: Maas, Matthijs, and Villalobos, José Jaime. 'International AI institutions: A literature review of models, examples, and proposals.' *Institute for Law & AI*, AI Foundations Report 1. (September 2023). https://www.legalpriorities.org/research/international-ai-institutions

[793] Ho, Lewis, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, et al. 'International Institutions for Advanced AI'. arXiv, 10 July 2023. https://doi.org/10.48550/arXiv.2307.04699.

International AI Organization (IAIO) to certify state jurisdictions for compliance with international AI oversight standards to enable states to prohibit the imports of goods "whose supply chains embody AI from non-IAIO-certified jurisdictions";[794] a proposal for an "international consortium" for evaluations of societal-scale risks from advanced AI;[795] a "Global Organization for High-Risk Artificial Intelligence (GOHAI)";[796] and many other proposals.[797]

[794] Trager, Robert, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, et al. 'International Governance of Civilian AI: A Jurisdictional Certification Approach'. arXiv, 29 August 2023. https://doi.org/10.48550/arXiv.2308.15514.

[795] Gruetzemacher, Ross, Alan Chan, Kevin Frazier, Christy Manning, Štěpán Los, James Fox, José Hernández-Orallo, et al. 'An International Consortium for Evaluations of Societal-Scale Risks from Advanced AI'. arXiv, 24 October 2023. https://doi.org/10.48550/arXiv.2310.14455.

[796] Llerena, Stephan. 'Global Governance of High-Risk Artificial Intelligence', 27 October 2023. (draft manuscript).

[797] See Maas, Matthijs, and Villalobos, José Jaime. 'International AI institutions: A literature review of models, examples, and proposals.' *Institute for Law & AI*, AI Foundations Report 1. (September 2023). https://www.legalpriorities.org/research/international-ai-institutions

Conclusion

The recent advances in AI have turned global public attention to this technology's capabilities, impacts, and risks. AI's significant present-day impacts and the prospect that these will only spread and scale further as these systems get increasingly advanced have firmly fixed this technology as a preeminent challenge for law and global governance this century.

In response, the disparate community of researchers that have explored aspects of these questions over the past years may increasingly be called upon to translate that research into rigorous, actionable, legitimate, and effective policies. They have developed—and continue to produce—a remarkably far-flung body of research, drawing on a diverse range of disciplines and methodologies. The urgency of action around advanced AI accordingly create a need for this field to increase the clarity of its work and its assumptions, to identify gaps in its approaches and methodologies where it can learn from yet more disciplines and communities, to improve coordination amongst lines of research, and to improve legibility of its argument and work to improve constructive scrutiny and evaluation of key arguments and proposed policies.

This review has not remotely achieved these goals—as no single document or review can. Yet by attempting to distill and disentangle key areas of scholarship, analysis, and policy advocacy, it hopes to help contribute to greater analytical and strategic clarity, more focused and productive research, and better-informed public debates and policymaker initiatives on the critical global challenges of advanced AI.

Also in this series

→ Maas, Matthijs, and Villalobos, José Jaime. 'International AI institutions: A literature review of models, examples, and proposals.' *Institute for Law & AI*, **AI Foundations Report 1**. (September 2023). https://www.law-ai.org/international-ai-institutions

→ Maas, Matthijs, 'AI is like… A literature review of AI metaphors and why they matter for policy.' *Institute for Law & AI*. **AI Foundations Report 2**. (October 2023). https://www.law-ai.org/ai-policy-metaphors

→ Maas, Matthijs, 'Concepts in advanced AI governance: A literature review of key terms and definitions.' *Institute for Law & AI*. **AI Foundations Report 3**. (October 2023). https://www.law-ai.org/advanced-ai-gov-concepts

INSTITUTE
FOR LAW & AI

law-ai.org