INSTITUTE
FOR LAW & AI

Concepts in advanced AI governance

A literature review of key terms and definitions

AI Foundations Report 3 | Matthijs Maas | October 2023

law-ai.org

Concepts in advanced AI governance

A literature review of key terms and definitions

Institute for Law & AI – AI Foundations Report 3

October 2023 | Matthijs Maas[1]

Abstract

As AI systems have become increasingly capable, policymakers, the public, and the field of AI governance have begun to consider the potential impacts and risks from these systems—and the question of how best to govern such increasingly advanced AI. Call this field "Advanced AI Governance." However, debates within and between these communities often lack clarity over key concepts and terms. In response, this report provides an overview, taxonomy, and preliminary analysis of many cornerstone ideas and concepts within advanced AI governance.

To do so, it first reviews three different purposes for seeking definitions (technological, sociotechnical, and regulatory), and discusses why and how terminology matters to both the study and the practice of AI governance. Next, the report surveys key definitions in advanced AI governance. It reviews 101 definitions across 69 terms that have been coined for advanced AI systems, within four categories: (1) essence-based concepts that focus on the anticipated *form* of advanced AI, (2) development-based terms that emphasize the hypothesized *pathways* towards advanced AI, (3) sociotechnical-change-based terms that center the *societal impacts* of such AI, and (4) risk-based terms that highlight specific *critical capabilities* of advanced AI systems. The report then reviews distinct definitions of the tools of (AI) "policy" and "governance", different paradigms within the field of advanced AI governance, and different concepts around theories of change. By disentangling these terms and definitions, this report aims to facilitate more productive conversations between AI researchers, academics, policymakers, and the public on the key challenges of advanced AI.

Cite as: 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

[1] I thank Shahar Avin, José Hernández-Orallo, Victoria Krakovna, Laurens Prins, Christoph Winter, Noam Kolt, Olle Häggström, Kevin Frazier, Lucy Lim, Marco Almada, John Danaher, Esben Kran, John Burden, Zach Stein-Perlman, John-Clark Levin, Carlos Ignacio Gutierrez, Lauro Langosco, Aishwarya Saxena, Gavin Leech, and one person who wished to remain anonymous. I especially thank Suzanne Van Arsdale, José Jaime Villalobos, and Kevin Frazier for in-depth substantive, structural and editorial insights that have greatly improved this document. I am also grateful to Daisy Newbold-Harrop, Erin Cooper, and Wes Cowley for editorial assistance in preparing this report for publication. All views expressed, and especially any remaining errors, are my own.

Executive Summary

This report provides an overview, taxonomy, and preliminary analysis of many cornerstone ideas and concepts in the emerging field of advanced AI governance.

Aim: The aim of this report is to contribute to improved analysis, debate, and policy by providing greater clarity around core terms and concepts. Any field of study or regulation can be improved by such clarity.

As such, this report reviews definitions for four categories of terms: the *object* of analysis (e.g., advanced AI), the *tools* for intervention (e.g., "governance" and "policy"), the reflexive definitions of the *field* of "advanced AI governance", and its *theories of change*.

Summary: In sum, this report:

I. Discusses three different purposes for seeking definitions for AI technology, discusses the importance of such terminology in shaping AI policy and law, and discusses potential criteria for evaluating and comparing such terms.

II. Reviews concepts for advanced AI, covering a total of 101 definitions across 69 terms, including terms focused on:
    1. the *forms* of advanced AI,
    2. the (hypothesized) *pathways* towards those advanced AI systems,
    3. the technology's large-scale *societal impacts*, and
    4. particular *critical capabilities* that advanced AI systems are expected to achieve or enable.

III. Reviews concepts within "AI governance", such as nine analytical terms used to define the tools for intervention (e.g., AI strategy, policy, and governance), four terms used to characterize different approaches within the field of study, and five terms used to describe theories of change.

The terms are summarized below in Table 1. Appendices provide detailed lists of definitions and sources for all the terms covered as well as a list of definitions for nine other auxiliary terms within the field.

Table 1: Taxonomy of surveyed terms

Category		**Surveyed terms**
Form of advanced AI (essence-based definitions)	Mind-like	→ Strong AI
	Autonomous	→ Autonomous machine/artificial intelligence → General artificial intelligence
	Human-like	→ Human-level AI (HLAI)
	General-purpose	→ Foundation models → General-purpose AI systems (GPAIS) → Comprehensive AI services (CAIS)
	General-purpose and human-level performance	→ Artificial general intelligence (AGI) *[task performance definition]* → Robust artificial intelligence
	General-purpose and beyond-human	→ AI+ → (Artificial) superintelligence (ASI)

	performance	→ Superhuman general-purpose AI (SGPAI) → Highly-capable foundation models
Pathways towards advanced AI (development-based definitions)	First-principles	→ De novo AGI
	Scaling	→ Prosaic AGI → Frontier (AI) model *[compute-threshold definition]*
	Evolutionary	→ [AGI] from evolution
	Reward-based	→ [AGI] from powerful reinforcement learning agents → Powerful deep learning models
	Bootstrapping	→ Seed AI
	Neuro-inspired	→ NeuroAI → Brain-like AGI → Neuromorphic AGI
	Neuro-emulated	→ Whole-brain-emulation (WBE) → Digital people *[emulation definition]*
	Neuro-integrationist	→ Brain-computer interfaces (BCI)
	Embodiment	→ Embodied agent
	Modular cognitive architecture	→ (N/A)
	Hybrid	→ Hybrid AI
Overall societal impacts of advanced AI (sociotechnical-change-based definitions)		→ (Strategic) general-purpose technology (GPT) → General-purpose military transformation (GMT) → Transformative AI (TAI) → Radically transformative AI (RTAI) → AGI *[economic competitiveness definition]* → Machine superintelligence *[form & impact definition]*
Critical capabilities of advanced AI (risk-based definitions)	Moral and philosophical	→ Artificial/Machine consciousness → Digital minds → Digital people *[capability definition]* → Sentient artificial intelligence → Robot rights catastrophe → (Negative) synthetic phenomenology → Suffering risks → Adversarial technological maturity
	Economic	→ High-level machine intelligence (HLMI) → Tech company singularity / fully general tech company → Artificial capable intelligence (ACI)
	Legal	→ Advanced artificial judicial intelligence (AAJI) → Technological-legal lock-in → Legal singularity

	Scientific	→ Process-automating science and technology (PASTA) → Scientist model
	Strategic and military	→ Decisive strategic advantage (DSA) → Singleton
	Political	→ Stable (digital) totalitarianism → Value lock-in → Actually existing AI (AEAI)
	Exponential	→ Intelligence explosion → Autonomous replication in the real world → Autonomous AI research → Duplicator
	Hazardous	→ Advanced AI → High-risk AI system → AI system of concern → Prepotent AI → APS system / Power-seeking AI → WIDGET → Rogue AI → Frontier (AI) model *[relative-capabilities-threshold definition]* → Frontier (AI) model *[dangerous-capabilities-threshold definition]* → Highly-capable systems of concern.
Tools for intervention	Strategy	→ AI strategy research → AI strategy → Long-term impact strategies → AI macrostrategy
	Policy	→ AI policy → AI policymaking strategy
	Governance	→ AI governance → Collaborative governance of AI technology → AGI safety and governance practices
Field (i.e., schools or paradigms of advanced AI governance)		→ (Advanced) AI governance → Transformative AI governance → Long-term AI governance → Longtermist AI governance
Theories of change (i.e., praxis)		→ (Analytic) frame → Theory of impact → Path to impact → Theory of change → Theory of victory

Table of Contents

Introduction

As AI systems have become increasingly capable and have had increasingly public impacts, the field that focuses on governing advanced AI systems has come into its own.

While researchers come to this issue with many different motivations, concerns, or hopes about AI—and indeed with many different perspectives on or expectations about the technology's future trajectory and impacts—there has grown an emerging field of researchers, policy practitioners, and activists concerned with and united by what they see as the increasingly significant and pivotal societal stakes of AI. Along with significant disagreements, many in this emerging community share the belief that shaping the transformative societal impacts of advanced AI systems is a top global priority.[2] However, this field still lacks clarity regarding not only many key empirical and strategic questions but also many key terms that are used.

Background: This lack of clarity matters because the recent wave of progress in AI, driven especially but not exclusively by the dramatic success of large language models (LLMs), has led to an accumulation of a wide range of new terms to describe these AI systems. Yet many of these terms—such as "foundation model",[3] "generative AI",[4] or "frontier AI"[5]—do not always have clear distinctions[6] and are often used interchangeably.[7] They moreover emerge on top of and alongside a wide range of past terms, concepts, and words that have been used in the past decades to refer to (potential) advanced AI systems, such as "strong AI", "artificial general intelligence", or "transformative AI". What are we to make of all of these terms?

[2] See Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, et al. 'Managing AI Risks in an Era of Rapid Progress', n.d. https://managing-ai-risks.com/.; Center for AI Safety. 'Statement on AI Risk', 30 May 2023. https://www.safe.ai/statement-on-ai-risk. And section III(2) below.

[3] "Foundation model" was originally defined as "models trained on broad data at scale [...] that are adaptable to a wide range of downstream tasks." See Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 'On the Opportunities and Risks of Foundation Models'. *arXiv:2108.07258 [Cs]*, 16 August 2021. http://arxiv.org/abs/2108.07258. See Jones, Elliot. 'Explainer: What Is a Foundation Model?' Ada Lovelace Institute, 17 July 2023. https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/. See also: Hausenloy, Jason, and Claire Dennis. 'Towards a UN Role in Governing Foundation Artificial Intelligence Models'. United Nations University - Centre for Policy Research, 19 July 2023. https://unu.edu/cpr/working-paper/towards-un-role-governing-foundation-artificial-intelligence-models. Pg. 8-9.

[4] "Generative AI" has been defined as "A type of AI system that can create a wide variety of data, such as images, videos, audio, text and 3D models" and "AI systems that can generate content based on user inputs such as text prompts [where] the content types (also known as modalities) that can be generated include like images, video, text and audio." Jones, Elliot. 'Explainer: What Is a Foundation Model?'. Alternately, it has been defined as: "models that input and output any combination of image, audio, video, and text. This includes transformer-based systems, such as large language models, diffusion-based systems, and hybrid architectures." See Weidinger, Laura, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, et al. 'Sociotechnical Safety Evaluation of Generative AI Systems'. arXiv, 18 October 2023. https://doi.org/10.48550/arXiv.2310.11986. Pg. 6.

[5] "Frontier AI" has been defined in various ways. For instance, as "large-scale machine-learning models that exceed the capabilities currently present in the most advanced existing models, and can perform a wide variety of tasks." 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/. But see also the other varying definitions of the term, discussed under Part II(2) and II(4), and in Appendices 1B and 1D.

[6] For comparisons and discussion of these overlapping terminologies, see Toner, Helen. 'What Are Generative AI, Large Language Models, and Foundation Models?' *Center for Security and Emerging Technology* (blog), 12 May 2023. https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/. See also 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. Pg. 3.

[7] Jones, Elliot. 'Explainer: What Is a Foundation Model?'

Rationale: Critically, debates over terminology in and for advanced AI are not just semantics—these terms matter. In a broad sense, framings, metaphors, analogies, and explicit definitions can strongly affect not just developmental pathways for technology but also policy agendas and the efficacy and enforceability of legal frameworks.[8] Indeed, different terms have already become core to major AI governance initiatives—with "general-purpose AI" serving as one cornerstone category in the EU AI Act[9] and "frontier AI models" anchoring the 2023 UK AI Safety Summit.[10] The varying definitions and implications of such terms may lead to increasing contestation,[11] as well they should: Extensive work over the past decade has shown how different terms for "AI" import different regulatory analogies[12] and have implications for crafting legislation.[13] We might expect the same to hold for the new generation of terms used to describe advanced AI and to center and focus its governance.[14]

Aim: The aim of this report is to contribute to improved analysis, debate, and policy by providing greater clarity around core terms and concepts. Any field of study or regulation can be improved by such clarity. Such literature reviews may not just contribute to a consolidation of academic work, but can also refine public and policy debates.[15] Ideally, they provide foundations for a more deliberate and reflexive choice over what concepts and terms to use (and which to discard), as well as a more productive refinement of the definition and/or operationalization of cornerstone terms.

Scope: In response, this report considers four *types* of terms, including potential concepts and definitions for each of the following:

[8] 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. (2023). https://www.law-ai.org/ai-policy-metaphors

[9] European Parliament. 'DRAFT Compromise Amendments on the Draft Report Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts', 9 May 2023. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf. As discussed in Jones, Elliot. 'Explainer: What is a Foundation Model?'.

[10] UK Government. 'AI Safety Summit: Introduction'. GOV.UK, 25 September 2023. https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html.

[11] See for example Henshall, Will. 'The Heated Debate Over Who Should Control Access to AI'. Time, 25 August 2023. https://time.com/6308604/meta-ai-access-open-source/.; Davies, Matt, and Michael Birtwistle. 'Seizing the "AI Moment": Making a Success of the AI Safety Summit'. Ada Lovelace Institute, 7 September 2023. https://www.adalovelaceinstitute.org/blog/ai-safety-summit/.

[12] See again Maas, Matthijs, 'AI is Like… A Literature Review of AI Metaphors and Why They Matter for Policy.' (2023). See also the discussion in Part I(2), below.

[13] Schuett, Jonas. 'Defining the Scope of AI Regulations'. *Law, Innovation and Technology* 15, no. 1 (3 March 2023): 1–23. https://doi.org/10.1080/17579961.2023.2184135.

[14] For a related, recent attempt to clarify and operationalize terminology around the term "AGI" specifically, see also: Morris, Meredith Ringel, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 'Levels of AGI: Operationalizing Progress on the Path to AGI'. arXiv, 4 November 2023. https://doi.org/10.48550/arXiv.2311.02462. For a related informal discussion of many of these terms, see: Guest, Oliver. 'What Term to Use for AI in Different Policy Contexts?' Effective Altruism Forum, 6 September 2023. https://forum.effectivealtruism.org/posts/9Y5YzNDMdYYg6hjwD/what-term-to-use-for-ai-in-different-policy-contexts. Another overview of some common terms is given in: Chapman, David. *Better without AI*, 2023. https://betterwithout.ai/. (discussing and critiquing the different concepts of "superintelligence", "mind-like AI", "autonomous AI agents", "AGI", "transformative AI"). For a related project that aims to collect a range of (legal) definitions for artificial-intelligence-related terms, see also: SAIDD. 'Statutory Artificial Intelligence Definitions Database'. SAIDD. Accessed 23 October 2023. https://www.saidd.info.

[15] See broadly: Clancy, Matt. 'Literature Reviews and Innovation'. Substack newsletter. *What's New Under the Sun* (blog), 2 October 2023. https://mattsclancy.substack.com/p/literature-reviews-and-innovation?post_id=137592816&r=4315a.

1. the core *objects of analysis*—and the targets for policy (i.e., what is the "advanced AI" to be governed?),
2. the *tools* for intervention to be used in response (i.e., what is the range of terms such as "policy", "governance", or "law"?),
3. the *field* or community (i.e., what are current and emerging accounts, projects, or approaches within the broader field of advanced AI governance?), and
4. the *theories of change* of this field (i.e., what is this field's praxis?).

Disclaimers: This project comes with some important caveats for readers.

First, this report aims to be relatively broad and inclusive of terms, framings, definitions, and analogies for (advanced) AI. In doing so, it draws from both older and recent work and from a range of sources from academic papers to white papers and technical reports to public fora.

Second, this report is primarily concerned with mapping the conceptual landscape and with understanding the (regulatory) implications of particular terms. As such, it is less focused on policing the appropriateness or coherence of particular terms or concepts. Consequently, with regard to advanced AI it covers many terms that are still highly debated or contested or for which the meaning is unsettled. Not all the terms covered are equally widely recognized, used, or even accepted as useful in the field of AI research or within the diverse fields of the AI ethics, policy, law, and governance space. Nonetheless, this report will include many of these terms on the grounds that a broad and inclusive approach to these concepts serves best to illuminate productive future debate. After all, even if some terms are (considered to be) "outdated," it is important to know where such terms and concepts have come from and how they have developed over time. If some terms are contested or considered "too vague," that should precisely speak in favor of aiming to clarify their usage and relation to other terms. This will either allow the (long overdue) refinement of concepts or will at least enable an improved understanding of when certain terms are not usefully recoverable. In both cases, it will facilitate greater clarity of communication.

Third, this review is a snapshot of the state of debate at one moment. It reviews a wide range of terms, many of which have been coined recently and only some of which may have staying power. This debate has developed significantly in the last few years and will likely continue to do so.

Fourth, this review will mostly focus on *analytical* definitions of or for advanced AI along four approaches.[16] In so doing, it will on this occasion mostly omit detailed exploration of a fifth, *normative* dimension to defining AI, which would focus on reviewing especially *desirable* types of advanced AI systems that (in the view of some) ought to be pursued or created. Such a review would cover a range of terms such as "ethical

[16] Along form, pathways, broad societal impacts, and critical capabilities. See Section II.

AI”,[17] “responsible AI”,[18] “explainable AI”,[19] “friendly AI”,[20] “aligned AI”,[21] “trustworthy AI”,[22] “provably-safe AI”,[23] “human-centered AI”,[24] “green AI”,[25] “cooperative AI”,[26] “rights-respecting AI”,[27] “predictable AI”,[28] “collective intelligence”,[29] and “digital plurality”,[30] amongst many other terms and concepts. At present, this report will not focus in depth on surveying these terms, since only some of them were articulated in the context of or in consideration of especially advanced AI systems. However, many or all of these terms are capability-agnostic and so could clearly be extended to or reformulated for more capable, impactful, or dangerous systems. Indeed, undertaking such a deepening and extension of the taxonomy

[17] See Jobin, Anna, Marcello Ienca, and Effy Vayena. ‘The Global Landscape of AI Ethics Guidelines’. *Nature Machine Intelligence*, 2 September 2019, 1–11. https://doi.org/10.1038/s42256-019-0088-2.

[18] Dignum, Virginia. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms. Cham: Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-30371-6.; and see also: Prabhakaran, Vinodkumar, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. ‘A Human Rights-Based Approach to Responsible AI’. arXiv, 6 October 2022. https://doi.org/10.48550/arXiv.2210.02667.

[19] See Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. ‘Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI’. *Information Fusion* 58 (1 June 2020): 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

[20] See Yudkowsky, Eliezer. ‘Friendly Artificial Intelligence’. In *Singularity Hypotheses*, edited by Amnon H. Eden, James H. Moor, Johnny H. Søraker, and Eric Steinhart, 181–95. The Frontiers Collection. Springer Berlin Heidelberg, 2012. https://doi.org/10.1007/978-3-642-32560-1_10.

[21] See 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; and also Hilton, Benjamin. ‘Preventing an AI-Related Catastrophe - Problem Profile’. 80,000 Hours, 25 August 2022. https://80000hours.org/problem-profiles/artificial-intelligence/. Ftn 29 (reviewing several different definitions of the term ‘ alignment’ used in this literature).

[22] Stix, Charlotte. ‘Artificial Intelligence by Any Other Name: A Brief History of the Conceptualization of “Trustworthy Artificial Intelligence”’. *Discover Artificial Intelligence* 2, no. 1 (21 December 2022): 26. https://doi.org/10.1007/s44163-022-00041-5; see also 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.; and Avin, Shahar, Haydn Belfield, Miles Brundage, Gretchen Krueger, Jasmine Wang, Adrian Weller, Markus Anderljung, et al. ‘Filling Gaps in Trustworthy Development of AI’. *Science* 374, no. 6573 (10 December 2021): 1327–29. https://doi.org/10.1126/science.abi7176.

[23] Tegmark, Max, and Steve Omohundro. ‘Provably Safe Systems: The Only Path to Controllable AGI’. arXiv, 4 September 2023. https://doi.org/10.48550/arXiv.2309.01933.

[24] Shneiderman, Ben. *Human-Centered AI*. Oxford, New York: Oxford University Press, 2022.

[25] Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni. ‘Green AI’. *arXiv:1907.10597 [Cs, Stat]*, 22 July 2019. http://arxiv.org/abs/1907.10597.

[26] 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.

[27] 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

[28] Zhou, Lexin, Pablo A. Moreno-Casares, Fernando Martínez-Plumed, John Burden, Ryan Burnell, Lucy Cheke, Cèsar Ferri, et al. ‘Predictable Artificial Intelligence’. arXiv, 9 October 2023. https://doi.org/10.48550/arXiv.2310.06167.

[29] The Collective Intelligence Project. ‘Whitepaper’. The Collective Intelligence Project, 2023. https://cip.org/whitepaper.

[30] 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. Pg. 10-12.

presented in this report in ways that engage more with the normative dimension of advanced AI would be very valuable future work.

Fifth, this report does not aim to definitively resolve debates—or to argue that all work should adopt one or another term over others. Different terms may work best in different contexts or for different purposes and for different actors. Indeed, given the range of actors interested in AI—whether from a technical engineering, sociotechnical, or regulatory perspective—it is not surprising that there are so many terms and such diversity in definitions even for single terms. Nonetheless, to be able to communicate effectively and learn from other fields, it helps to gain greater clarity and precision in the terms we use, whether these are terms referring to our objects of analysis, our own field and community, or our theory of action. Of course, achieving clarity on terminology is not itself sufficient. Few problems, technical or social or legal, may be solved exclusively by haggling over words. Nonetheless, a shared understanding facilitates problem solving. The point here is not to achieve full or definitive consensus but to understand disagreements and assumptions. As such, this report seeks to provide background on many terms, explore how they have been used, and consider the suitability of these terms for the field.[31] In doing so, this report highlights the diversity of terms in current use and provides context for more informed future study and policymaking.

Structure: Accordingly, this report now proceeds as follows.

Part I provides a background to this review by discussing three purposes to defining key terms such as AI. It also discusses why the choice for one or another term matters significantly from the perspective of AI policy and regulation, and finally discusses some criteria by which to evaluate the suitability of various terms and definitions for the specific purpose of regulation.

In Part II, this report reviews a wide range of terms for "advanced AI", across different approaches which variably focus on (a) the anticipated *forms* or design of advanced AI systems, (b) the hypothesized scientific *pathways* towards these systems, (c) the technology's broad *societal impacts*, or (d) the specific *critical capabilities* particular advanced AI systems are expected to achieve.

Part III turns from the object of analysis to the field and *epistemic community* of advanced AI governance itself. It briefly reviews three categories of concepts of use for understanding this field. First, it surveys different terms used to describe AI "strategy", "policy", or "governance" as this community understands the *available tools for intervention* in shaping advanced AI development. It then reviews different paradigms within the field of advanced AI governance as ways in which different voices within it have defined that *field*. Finally, it briefly reviews recent definitions for *theories of change* that aim to compare and prioritize interventions into AI governance.

Finally, three appendices list in detail all the terms and definitions offered, with sources, and offer a list of auxiliary definitions that can aid future work in this emerging field.[32]

[31] This also can ground research into high-level approaches to governing advanced AI systems. See informally Maas, Matthijs M. 'Strategic Perspectives on Transformative AI Governance: Introduction'. EA Forum, 2 July 2022. https://forum.effectivealtruism.org/posts/isTXkKprgHh5j8WQr/strategic-perspectives-on-transformative-ai-governance.

[32] These appendices may be helpful for readers to explore work in this field in more detail; to understand the longer history and evolution of many terms; and to consider the strengths and drawbacks of particular terms, and of specific language, for use in public debate, policy formulation, or even in direct legislative texts.

I. Defining "advanced AI (governance)": Background

Any quest for clarifying definitions of "advanced AI" is complicated by the already long-running, undecided debates over how to even define the more basic terms "AI" or, indeed, "intelligence".[33]

To properly evaluate and understand the relevance of different terms for AI, it is useful to first set out some background. In the first place, one should start by considering the purposes for which the definition is sought. *Why* or *how* do we seek definitions of "(advanced) AI"?

1. Three purposes for definitions

For instance, rather than trying to consider a universally best definition for AI, a more appropriate approach is to consider the implications of different definitions, or—to invert the question—to ask for what purpose we seek to define AI. We can consider (at least) three different rationales for defining a term like "AI":

1. **To build it (the technological research purpose):** In the first place, AI researchers or scientists may pursue definitions of (advanced) AI by defining it from the "inside," as a science.[34] The aim of such technical definitions of AI[35] is to clarify or create research-community consensus about (1) the range and disciplinary boundaries of the field—that is, what research programs and what computational techniques[36] count as "AI research" (both internally and externally to research funders or users); (2) the long-range goals of the field (i.e., the technical forms of advanced AI); and/or (3) the intermediate steps the field should take or pursue (i.e., the likely pathways towards such AI). Accordingly, this

[33] Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River: Pearson, 2016. Pg. 2. (providing a classic taxonomy of the ways in which AI researchers have defined "intelligence", distinguishing between systems that achieve "thinking humanly," "thinking rationally," "acting humanly," or "acting rationally"). See also Monett, Dagmar, Colin W. P. Lewis, Kristinn R. Thórisson, Joscha Bach, Gianluca Baldassarre, Giovanni Granato, Istvan S. N. Berkeley, et al. 'Special Issue "On Defining Artificial Intelligence"—Commentaries and Author's Response'. *Journal of Artificial General Intelligence* 11, no. 2 (1 February 2020): 1–100. https://doi.org/10.2478/jagi-2020-0003. Pg. 1.

[34] That is not to say all agree that a single definition is needed. Indeed, in the past some AI researchers themselves have been happy to shelve definitional questions, and "get on with it." Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, et al. 'Artificial Intelligence and Life in 2030'. One Hundred Years of Artificial Intelligence. Stanford, CA: Stanford University, September 2016. http://ai100.stanford.edu/2016-report. Pg. 12.

[35] For instance, Nilsson has defined the field of AI as being concerned with "making machines intelligent, [where] intelligence is that quality that enables an entity to function appropriately and with foresight in its environment." Nilsson, Nils J. *The Quest for Artificial Intelligence*. 1 edition. Cambridge ; New York: Cambridge University Press, 2009. Pg. xiii. Another broad and encapsulating "scientific" definition for AI has stated: "AI is a branch of computer science (CS), which is the scientific study of what problems can be solved, what tasks can be accomplished, and what features of the world can be understood computationally (i.e., using the language of Turing Machines), and then to provide algorithms to show how this can be done efficiently, practically, physically, and ethically. [...] Given that CS's primary question is "What is computable?", I take the focus of AI to be on whether cognition is computable." Rapaport, William J. 'What Is Artificial Intelligence?' *Journal of Artificial General Intelligence*, Special Issue "On Defining Artificial Intelligence"—Commentaries and Author's Response, 11, no. 2 (1 February 2020): 52–56. https://doi.org/10.2478/jagi-2020-0003.

[36] "Techniques" encompass a diverse and ever-evolving range of paradigms and approaches. Though for an older (2020) mapping, see for instance Hernandez-Orallo, Jose, Fernando Martınez-Plumed, Shahar Avin, Jess Whittlestone, and Seán Ó hÉigeartaigh. 'AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues'. In *European Conference on Artificial Intelligence*, 8, 2020. https://ecai2020.eu/papers/1364_paper.pdf. Pg. 3. (identifying 14 categories of AI techniques, with distinct subcategories and techniques).

definitional purpose aligns particularly closely with essence-based definitions (see Part II.1) and/or development-based definitions (see Part II.2) of advanced AI.

2. **To study it (the sociotechnical research purpose):** In the second place, experts (in AI, but especially in other fields) may seek to primarily understand AI's impacts on the world. In doing so, they may aim to define AI from the "outside," as a sociotechnical system including its developers and maintainers.[37] Such definitions or terms can aid researchers (or governments) who seek to understand the societal impacts and effects of this technology in order to diagnose or analyze the potential dynamics of AI development, diffusion, and application, as well as the long-term sociopolitical problems and opportunities. For instance, under this purpose researchers may aim to get to terms with understanding issues such as (1) (the geopolitics or political economy of) key AI inputs (e.g., compute, data, and labor), (2) how different AI capabilities[38] give rise to a spectrum of useful applications[39] in diverse domains, and (3) how these applications in turn produce or support new behaviors and societal impacts.[40] Accordingly, this purpose is generally better served by sociotechnical definitions of AI systems' impacts (see Part II.3) or risk-based definitions (see Part II.4).
3. **To regulate it (the regulatory purpose)**: Finally, regulators or academics motivated by appropriately regulating AI—either to seize the benefits or to mitigate adverse impacts—can seek to pragmatically delineate and define (advanced) AI as a legislative and regulatory target. In this approach, definitions of AI are to serve as useful handles for law, regulation, or governance.[41] In principle, this purpose can be well served by many of the definitional approaches: highly technology-specific regulations for instance can gain from focusing on development-based definitions of (advanced) AI. However, in practice regulation and governance is usually better served by focusing on the sociotechnical impacts or capabilities of AI systems.

Since it is focused on the field of "advanced AI governance," this report will primarily focus on the second and third of these purposes. However, it is useful to keep all three in mind.

[37] This discussion draws on: 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. 36–39).

[38] "Capabilities" are high-level abilities of AI systems that (as distinct from applications) are applicable across a range of datasets or domains. Such capabilities can therefore include different narrow but domain-agnostic functions that are of use in diverse contexts, or increasingly general capabilities that allow a system to perform well in diverse tasks (i.e., to become "general purpose"). Examples of such capabilities can include data classification, data generation, anomaly or pattern detection, prediction, optimization of complex systems and tasks, or autonomous operation of cyber-physical platforms or robots, amongst many others. However, there are many different taxonomies of such capabilities. See for instance the taxonomy of (capability) milestones presented by 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 Pg. 105. Or see more generally the characterization of an intelligent system in: Molina, Martin. 'What Is an Intelligent System?' arXiv, 18 December 2022. https://doi.org/10.48550/arXiv.2009.09083.

[39] "Applications" are an extremely diverse class of use cases. It is where AI techniques that enable certain capabilities refract into the full range of specific "useful" tasks that can be carried out for different principals: from email spam detection to facial recognition, from self-driving cars to energy grid optimization, from chatbots to deepfakcs, and from cybersecurity to lethal autonomous weapons systems, amongst others.

[40] See also Section II.3 (on definitions of advanced AI that focus on sociotechnical impacts).

[41] See also Veale, Michael, Kira Matus, and Robert Gorwa. 'AI and Global Governance: Modalities, Rationales, Tensions'. *Annual Review of Law and Social Science* 19, no. 1 (2023): null. https://doi.org/10.1146/annurev-lawsocsci-020223-040749. Pg. 3 (discussing different "conceptual targets" of AI regulation by comparing rules by the different aspects of practical AI that those rules seek to shape—whether the development of AI, its use, or its underlying infrastructures). I thank Marco Almada for this suggestion.

2. Why terminology matters to AI governance

Whether taking a sociotechnical perspective on the societal impacts of advanced AI or a regulatory perspective on adequately governing it, the need to pick suitable concepts and terms becomes acutely clear. Significantly, the implications and connotations of key terms matter greatly for law, policy, and governance. This is because, as reviewed in a companion report,[42] distinct or competing terms for AI—with their meanings and connotations—can influence all stages of the cycle from a technology's development to its regulation. They do so in both a broad and a narrow sense.

In the broad and preceding sense, the choice of term and definition can, explicitly or implicitly, import particular analogies or metaphors into policy debates that can strongly shape the direction—and efficacy—of the resulting policy efforts.[43] These framing effects can occur even if one tries to avoid explicit analogies between AI and other technologies, since apparently "neutral" definitions of AI still focus on one or another of the technology's "features" as the most relevant, framing policymaker perceptions and responses in ways that are not neutral, natural, or obvious. For instance, Murdick and others found that the particular definition one uses for what counts as "AI" research directly affects which (industrial or academic) metrics are used to evaluate different states' or labs' relative achievements or competitiveness in developing the technology—framing downstream evaluations of which nation is "ahead" in AI.[44] Likewise, Kraftt and colleagues found that whereas definitions of AI that emphasize "technical functionality" are more widespread among AI researchers, definitions that emphasize "human-like performance" are more prevalent among policymakers, which they suggest might prime policymaking towards future threats.[45]

Beyond the broad policy-framing impacts of technology metaphors and analogies, there is also a narrower sense in which terms matter. Specifically, within regulation, legislative and statutory definitions delineate the scope of a law and of the agency authorization to implement or enforce it[46]—such that the choice for a particular term for (advanced) AI may make or break the resulting legal regime.

Generally, within legislative texts, the inclusion of particular statutory definitions can play both *communicative* roles (clarifying legislative intent), and *performative* roles (investing groups or individuals with rights or obligations).[47] More practically, one can find different types of definitions that play distinct roles within regulation: (1) *delimiting* definitions establish the limits or boundaries on an otherwise ordinary meaning of a term, (2) *extending* definitions broaden a term's meaning to expressly include elements or components that might not normally be included in its ordinary meaning, (3) *narrowing* definitions aim to set limits or expressly

[42] 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.law-ai.org/ai-policy-metaphors

[43] Ibid. pg. 11-13.

[44] Murdick, Dewey, James Dunham, and Jennifer Melot. 'AI Definitions Affect Policymaking'. Center for Security and Emerging Technology, 2 June 2020. https://cset.georgetown.edu/research/ai-definitions-affect-policymaking/. (noting that "the competitive landscape varies significantly in sub-areas such as computer vision (where China leads), robotics (where China has made significant progress), and natural language processing (where the United States maintains its lead).", at 2).

[45] 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.

[46] Schuett, Jonas. 'Defining the Scope of AI Regulations'. *Law, Innovation and Technology* 15, no. 1 (3 March 2023): 1–23. https://doi.org/10.1080/17579961.2023.2184135.

[47] Price, Jeanne. 'Wagging, Not Barking: Statutory Definitions'. *Cleveland State Law Review* 60, no. 60 (2013): 999–1055. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2288824

exclude particular understandings, and (4) *mixed* definitions use several of these approaches to clarify components.[48]

Likewise, in the context of AI law, legislative definitions for key terms such as “AI” obviously affect the material scope of the resulting regulations.[49] Indeed, the effects of particular definitions have impacts on regulation not only ex ante, but also ex post: in many jurisdictions, legal terms are interpreted and applied by courts based on their widely shared “ordinary meaning.”[50] This means, for instance, that regulations that refer to terms such as “advanced AI”, “frontier AI”, or “transformative AI” might not necessarily be interpreted or applied in ways that are in line with how the term is understood within expert communities. All of this underscores the importance of our choice of terms—from broad and indirect metaphors to concrete and specific legislative definitions—when grappling with the impacts of this technology on society.

Indeed, the strong legal effects of different terms mean that there can be challenges for a law when it depends on a poorly or suboptimally specified regulatory term for the forms, types, or risks from AI that the legislation means to address. This creates twin challenges. On the one hand, picking suitable concepts or categories can be difficult at an early stage of a technology’s development and deployment, when its impacts and limits are not always fully understood—the so-called Collingridge dilemma.[51]

At the same time, the cost of picking and locking in the wrong terms within legislative texts can be significant. Beyond the opportunity costs, unreflexively establishing legal definitions for key terms can create the risk of downstream or later “governance misspecification.”[52]

Such governance misspecification may occur when regulation is originally targeted at a particular artifact or (technological) practice through a particular material scope and definition for those objects. The implicit assumption here is that the term in question is a meaningful proxy for the underlying societal or legal goals to

[48] Government of Canada, Department of Justice. ‘Legistics - Definitions’, 2 December 1999. https://www.justice.gc.ca/eng/rp-pr/csj-sjc/legis-redact/legistics/p1p5.html. I thank Suzanne Van Arsdale and Kevin Frazier for highlighting this taxonomy.

[49] Schuett, Jonas. ‘Defining the Scope of AI Regulations’. *Law, Innovation and Technology* 15, no. 1 (3 March 2023): 1–23. https://doi.org/10.1080/17579961.2023.2184135.

[50] Martínez, Eric, and Christoph Winter. ‘Ordinary Meaning of Existential Risk. LPP Working Paper No. 7-2022, 15 December 2022. https://doi.org/10.2139/ssrn.4304670. Others have suggested that courts will interpret definitions in ways that align with the median public opinion. See Dorf, Michael C. ‘Majoritarian Difficulty and Theories of Constitutional Decision Making’. *Journal of Constitutional Law* 13, no. 2 (2010): 283–304. Note that in certain circumstances, a court may refer to a technical meaning of a term to resolve ambiguity. See for instance Sullivan, Ruth. ‘Technical Meaning and Meanings Fixed by Law’. In *Statutory Interpretation*, 73–95. Irwin Law, 2016. However, even if a technical definition is invoked, this may not always be an easy resolution if there are many competing or overlapping technical definitions for the same term. In some contexts such as the US, where the meaning of a word is ambiguous, courts may apply a series of additional substantive canons of interpretation. See Baude, William, and Ryan D. Doerfler. ‘The (Not So) Plain Meaning Rule’. *The University of Chicago Law Review* 84, no. 2 (Spring 2017): 539–66. https://lawreview.uchicago.edu/print-archive/not-so-plain-meaning-rule. I thank Suzanne Van Arsdale for this suggestion; and both Suzanne and Kevin Frazier for work in this space.

[51] See Collingridge, David. *The Social Control of Technology*. New York: Palgrave Macmillan, 1981. See also: 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. Pg. 132-135. Nonetheless, there are also arguments in favor of the general feasibility of forward-looking, “anticipatory” regulation, even at an early stage. See Guston, David H. ‘Understanding “Anticipatory Governance”’. *Social Studies of Science* 44, no. 2 (April 2014): 218–42. https://doi.org/10.1177/0306312713508669. Armstrong, Harry, and Jen Rae. ‘A Working Model for Anticipatory Regulation’. Nesta, 2017. https://media.nesta.org.uk/documents/working_model_for_anticipatory_regulation_0_TpDHt7z.pdf.

[52] I thank Christoph Winter for introducing the term and concept.

be regulated. While that assumption may be appropriate and correct in many cases, there is a risk that if that assumption is wrong—either because of an initial misapprehension of the technology or because subsequent technological developments lead to that proxy term diverging from the legislative goals—the resulting technology law will less efficient, ineffective, or even counterproductive to its purposes.[53]

Such cases of governance misspecification can be seen in various cases of technology governance and regulation. For instance:

→ **The "high-performance computer" threshold in US 1990s export control regimes:** In the 1990s, the US established a series of export control regimes under the Wassenaar Arrangement, which set an initial threshold for "high-performance computers" at just 195 million theoretical operations per second (MTOPS); in doing so, the regime treated that technology as far too static and could not keep pace with Moore's Law.[54] As a result, the threshold had to be updated six times within a decade,[55] even as the regime became increasingly ineffective at preventing or even inhibiting US adversaries from accessing as much computing power as they needed, and it may even have become harmful to national security as it inhibited the domestic US tech industry.[56]

→ **The "in orbit" provision in the Outer Space Treaty:** In the late 1960s, the Outer Space Treaty aimed to outlaw positioning weapons of mass destruction in space. It therefore (as proxy) specified a ban on placing these weapons "in orbit."[57] This definition meant that there was a loophole to be exploited by the Soviet development of fractional orbital bombardment systems (FOBS), which were able to position nuclear weapons in space (on non-ballistic trajectories) without, strictly, putting them "in orbit."[58]

→ **Under- and overinclusive 2010s regulations on drones and self-driving cars:** Calo has chronicled how, in the early 2010s, various US regulatory responses to drones or self-driving cars defined these

[53] In a legal context, this echoes HLA Hart's classic "no vehicles in the park" dilemma–the situation where a certain rule (say, at a city park) was originally formulated to ban certain objects (e.g., motor vehicles) from a park, but where it was phrased without awareness of other objects (e.g., bicycles, roller skates, electric wheelchairs, and drones) that might fall under this terminology, creating later uncertainty over whether it would—or why it should—apply to these new objects. See Hart, H. L. A. 'Positivism and the Separation of Law and Morals'. *Harvard Law Review* 71, no. 4 (February 1958): 593. https://doi.org/10.2307/1338225. Pg. 607. See also Schlag, Pierre. 'No Vehicles in the Park'. *Seattle University Law Review* 23 (1999): 381–89. https://digitalcommons.law.seattleu.edu/cgi/viewcontent.cgi?article=1623&context=sulr .

[54] Picker, Colin B. 'A View from 40,000 Feet: International Law and the Invisible Hand of Technology'. *Cardozo Law Review* 23 (2001): 151–219. Pg. 212.

[55] Ibid. pg. 212-213. ("for civilian end-users, the Tier 3 computer export control threshold in 1991 was established at 195 MTOPS, and was increased in 1994 to 1,500 MTOPS; in 1996 to 7,000 MTOPS; in August 1999 to 12,300 MTOPS; in February 2000 to 20,000 MTOPS; in August 2000 to 28,000 MTOPS; and in January 2001 to 85,000 MTOPS.").

[56] ibid.

[57] Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies, 610 UNTS 205 § (1967). Article IV.

[58] Maas, 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf , pg. 197-205. Citing Garthoff, Raymond L. 'Banning the Bomb in Outer Space'. *International Security* 5, no. 3 (1980): 25–40. https://doi.org/10.2307/2538418.; Deudney, Daniel. *Dark Skies: Space Expansionism, Planetary Geopolitics, and the Ends of Humanity*. Oxford, New York: Oxford University Press, 2020. Pg. 413. The incident caused a political uproar in the US, and no further tests of the system were conducted, although the launchers stayed operational. Subsequently, FOBS-type systems were explicitly prohibited by the SALT II agreement of 1979; while the US Senate did not ratify SALT II, the Soviet Union did comply with its terms, decommissioning or converting the remaining FOBS launchers by 1983. Gyűrösi, Miroslav. 'The Soviet Fractional Orbital Bombardment System Program'. Air Power Australia, 2 January 2010. http://www.ausairpower.net/APA-Sov-FOBS-Program.html.

technologies in ways that were either under- or overinclusive, leading to inefficiency or the repeal of laws.[59]

Thus, getting greater clarity in our concepts and terminology for advanced AI will be critical in crafting effective, resilient regulatory responses—and in avoiding brittle missteps that are easily misspecified.

Given all the above, the aim in this report is not to find the "correct" definition or frame for advanced AI. Rather, it considers that different frames and definitions can be more useful for specific purposes or for particular actors and/or (regulatory) agencies. In that light, we can explore a series of broad starting questions, such as:

1. What different definitions have been proposed for advanced AI? What other terms could we choose?
2. What aspects of advanced AI (e.g., its form and design, the expected scientific principles of its development pathways, its societal impacts, or its critical capabilities) do these different terms focus on?
3. What are the regulatory implications of different definitions?

In sum, this report is premised on the idea that exploring definitions of AI (and related terms) matters, whether we are trying to understand AI, understand its impacts, or govern them effectively.

3. Criteria for definitions

Finally, we have the question of how to formulate relevant criteria for suitable terms and definitions for advanced AI. In the first place, as discussed above, this depends on one's definitional purpose.

Nonetheless, from the specific perspective of regulation and policymaking, what are some good criteria for evaluating suitable and operable definitions for advanced AI? Notably, Jonas Schuett has previously explored legal approaches to defining the basic term "AI". He emphasizes that to be suitable for the purpose of governance, the choice of terms for AI should meet a series of requirements for all good legal definitions—namely that terms are neither (1) overinclusive nor (2) underinclusive and that they are (3) precise, (4) understandable, (5) practicable, and (6) flexible.[60] Other criteria have been proposed: for instance, it has been suggested that an additional desiderata for a useful regulatory definition for advanced AI might include something like ex ante clarity—in the sense that the definition should allow one to assess, for a given AI model, whether it will meet the criteria for that definition (i.e., whether it will be regulated within some regime), and ideally allow this to be assessed in advance of deployment (or even development) of that model.[61]

[59] See Calo, Ryan. 'The Case for a Federal Robotics Commission'. Brookings Institute Center for Technology Innovation, 1 September 2014. https://papers.ssrn.com/abstract=2529151. Pg. 6, 8 (discussing a 2011 incident where Nevada passed accidentally overinclusive self-driving car regulations, which had to be repealed after it turned out that they inadvertently imposed stringent obligations on existing vehicles with partially-autonomous features, as well as cases where US laws against drone surveillance ended up focusing far too much on flying drones rather than other mobile robots).

[60] Schuett, Jonas. 'Defining the Scope of AI Regulations'. *Law, Innovation and Technology* 15, no. 1 (3 March 2023): 1–23. https://doi.org/10.1080/17579961.2023.2184135. pg. 5–6.

[61] 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'. Pg. 34. (arguing that a good definition "should limit its scope to only those models for which there is good reason to believe they have sufficiently dangerous capabilities [...and moreover] it should be possible to determine whether a planned model will be regulated ex ante, before the model is developed. For example, the definition could be based on the model development process that will be used (e.g., data,

Certainly, these criteria remain contested and are likely incomplete. In addition, there may be trade-offs between the criteria, such that even if they are individually acceptable, one must still strike a workable balance between them.[62]

II. Defining the object of analysis: Terms for advanced AI

Having briefly discussed the different definitional purposes, the relevance of terms for regulation, and potential criteria for evaluating definitions, this report now turns to survey the actual terminology for advanced AI.

Within the literature and public debate, there are many terms used to refer to the conceptual cluster of AI systems that are advanced—i.e., that are sophisticated and/or are highly capable and/or could have transformative impacts on society.[63] However, because of this diversity of terms, not all have featured equally strongly in governance or policy discussions. To understand and situate these terms, it is useful to compare their definitions with others and to review different approaches to defining advanced AI.

In Schuett's model for "legal" definitions for AI, he has distinguished four types of definitions, which focus variably on (1) the *overarching term* "AI", (2) particular *technical approaches* in machine learning, (3) *specific applications* of AI, and (4) *specific capabilities* of AI systems (e.g., physical interaction, ability to make automated decisions, ability to make legally significant decisions).[64]

Drawing on Schuett's framework, this report draws a similar taxonomy for common definitions for advanced AI. In doing so, it compares between different approaches that focus on one of four features or aspects of advanced AI.

1. The anticipated technical form or design of AI systems (essence-based approaches);
2. The proposed scientific pathways and paradigms towards creating advanced AI (development-based approaches);
3. The broad societal impacts of AI systems, whatever their cognitive abilities (sociotechnical-change-based approach);

algorithms, and compute), rather than relying on ex post features of the completed model (e.g., capabilities, performance on evaluations).")

[62] I thank Marco Almada for this observation.

[63] For a good discussion of some terminology, and especially the distinction between the key terms "artificial general intelligence" (AGI) and "transformative AI" (TAI), see: Gruetzemacher, Ross, and Jess Whittlestone. 'The Transformative Potential of Artificial Intelligence'. *Futures* 135 (2022): 102884. https://doi.org/10.1016/j.futures.2021.102884. See also later discussion in this section and in Appendix 1. For a broader recent survey of terms common in these debates, see also: Jones, Elliot. 'Explainer: What Is a Foundation Model?' Ada Lovelace Institute, 17 July 2023. https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

[64] Ibid. 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. 36–39). For a discussion of different models used to classify AI systems in the context of operationalizing AI ethics, see: Mökander, Jakob, Margi Sheth, David S. Watson, and Luciano Floridi. 'The Switch, the Ladder, and the Matrix: Models for Classifying AI Systems'. *Minds and Machines*, 4 January 2023. https://doi.org/10.1007/s11023-022-09620-y.

4. The specific critical capabilities[65] that could potentially enable extreme impacts in particular domains (risk-based approaches).

Each of these approaches has a different focus, object, and motivating question (Table 2).

Table 2: Overview of approaches to defining advanced AI

	Focus	**Approach**	**Object of definition**	**Motivating question**
1	Form and architecture of advanced AI	Essence-based	Individual AI systems	What is an advanced AI?
2	Pathways towards advanced AI	Development-based	Individual AI systems produced through a particular technique or architecture	How could we build advanced AI?
3	General societal impacts of advanced AI	Sociotechnical-change-based	Aggregate effects of many AI systems	What are the societal impacts of advanced AI?
4	Critical capabilities of particular advanced AI systems	Risk-based	Capabilities achieved by particular AI systems	What are key risks from advanced AI?

This report will now review these categories of approaches in turn. For each, it will broadly (1) discuss that approach's core definitional focus and background, (2) list the terms and concepts that are characteristic of it, (3) provide some brief discussion of common themes and patterns in definitions given to these terms,[66] and (4)

[65] Note, this taxonomy is not perfect, as it leaves two ambiguities. In the first place, one can ask what fully distinguishes advanced AI definitions under approach 1 (forms) from those in approach 4 (critical capabilities)? After all, both involve terms that center the traits or capabilities of (particular) AI systems—often as measured by their ability to pass one or more benchmarks or evaluation tests (or inflict particular types of harm). Nonetheless, this report retains the distinction as useful, considering "forms" to reflect broader general capabilities and properties of systems, and "critical capabilities" as focusing on particular skills these systems can display in particular domains that suffice for them to have significant impacts or manifest important risks, even if they still lack full generality or many other skills or traits in other domains. This distinction relevantly carves the field at its joints, as it distinguishes the often-separate bodies of work that have particularly focused on (and often but not always been optimistic about) the creation of advanced AI systems from those that may be more reserved or even concerned and that express particular concerns over what advanced AI systems might do with particular capabilities. I thank Suzanne Van Arsdale for prompting this observation. In the second place, one can ask—what fully distinguishes definitions under approach 3 (societal impacts) from those in approach 4 (critical capabilities)? There is again some ambiguity here, since some terms that are here categorized as "critical capabilities" (such as "value lock-in" or "singleton") could also easily be described as societal impacts (and indeed, are critical through virtue of the significant impacts on society that they enable). To clarify, I consider terms that focus on societal impacts to be distinct because they (a) focus on the aggregate societal effects of many systems and can often be relatively agnostic over the precise forms, capabilities, or traits of individual systems; and (b) often focus on understanding broad sociotechnical impacts of advanced AI rather than specific risky outcomes of particular capabilities (to be avoided). However, this distinction is admittedly leaky.

[66] For the sake of brevity, the definitions for each term or concept are listed with sources in Appendix 1.

then provide some preliminary reflections on the suitability of particular terms within this approach, as well as of the approach as a whole, to provide usable analytical or regulatory definitions for the field of advanced AI governance.[67]

1. Essence-based definitions: Forms of advanced AI

Focus of approach: Classically, many definitions of advanced AI focus on the anticipated form, architecture, or design of future advanced AI systems.[68] These definitions as such focus on AI systems that instantiate particular forms of advanced intelligence,[69] for instance by instantiating an "actual mind" (that "really thinks"); by displaying a degree of autonomy; or by being human-like, general-purpose, or both in the ability to think, reason, or achieve goals across domains (see Table 3).

Terms: The form-centric approach to defining advanced AI accordingly encompasses a variety of terms, including strong AI, autonomous machine (/ artificial) intelligence, general artificial intelligence, human-level AI, foundation model, general-purpose AI system, comprehensive AI services, artificial general intelligence, robust artificial intelligence, AI+, (machine/artificial) superintelligence, superhuman general-purpose AI, and highly-capable foundation models.[70]

Definitions and themes: While many of these terms are subject to a wide range of different definitions (see Appendix 1A), they combine a range of common themes or patterns (see Table 3).

[67] That is, assuming a definitional purpose that is sociotechnical (#2) or regulatory (#3).

[68] See also: Gruetzemacher, Ross, and Jess Whittlestone. 'The Transformative Potential of Artificial Intelligence'. *Futures* 135 (2022): 102884. https://doi.org/10.1016/j.futures.2021.102884. Pg. 2 ("it is plausible that more advanced AI systems could precipitate dramatic societal changes. [...] Several different terms have been used to refer to the possibility of [...] humanlike AI systems with the potential to lead to such changes, [...] These notions all imply that most of our concern should be afforded to systems which are human-like or sufficiently general in their capabilities.").

[69] For another mapping of kinds of intelligent systems, see: Bhatnagar, Sankalp, Anna Alexandrova, Shahar Avin, Stephen Cave, Lucy Cheke, Matthew Crosby, Jan Feyereisl, et al. 'Mapping Intelligence: Requirements and Possibilities'. In *Philosophy and Theory of Artificial Intelligence 2017*, edited by Vincent C. Müller, 117–35. Studies in Applied Philosophy, Epistemology and Rational Ethics. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-96448-5_13.

[70] For different definitions for each of these terms with sources, see Appendix 1A.

Table 3: Form-focused definitions of advanced AI

Emphasis of definitional approach[71] Advanced AI is […]	**Term** [# of definitions surveyed][72]	**Common themes and patterns in definitions**
Mind-like	Strong AI [3]	→ Is a "mind" that "actually thinks"
Autonomous	Autonomous (machine / artificial) intelligence [2]	→ Learns more like animals and humans → Can adapt to external environmental challenges → Behavior driven by intrinsic objectives rather than by hard-wired programs
	General artificial intelligence [1]	→ Functions autonomously in novel circumstances
Human-like	Human-level AI (HLAI) [4]	→ Operates in a commonsense information environment → Able to do many of the things humans are able to do.
General-purpose	Foundation model [2]	→ Adaptive to many downstream tasks → Basis for other roles
	General-purpose AI systems (GPAIS) [4]	→ Can be adapted to a wide range of applications → Can be used for tasks for which it was not intentionally, specifically designed or trained
	Comprehensive AI services (CAIS) [1]	→ Recursive improvement of AI technologies in distributed systems, rather than unitary agents → Ecosystem of comprehensive superintelligent-level AI services, where agency is optional
General-purpose and human-level performance	Artificial general intelligence (AGI) *[task performance definitions]* [20]	→ Exhibits the broad range of general intelligence found in humans → Able to reason across a wide range of domains → Ability to develop a world model that is more accurate than that of humans
	Robust artificial intelligence [1]	→ Systematically and reliably applies its knowledge to a wide range of problems

[71] Note, this categorization is oversimplifying, since many of these terms also include some emphasis on the other traits.

[72] For the specific definitions for each of these terms, see Appendix 1A.

		→ Reasons flexibly and dynamically about the world
General-purpose and beyond-human performance	AI+ [1]	→ AI that is more intelligent than the most intelligent human
	(Machine / artificial) superintelligence (ASI) [7]	→ A(G)I that exceeds the best human performance in all domains
	Superhuman general-purpose AI (SGPAI) [1]	→ General-purpose AI (GPAI) that is simultaneously as good as or better than humans across nearly all tasks
	Highly-capable foundation models [1]	→ Foundation models that exhibit high performance across a broad domain, often performing as well as or better than a human

Suitability of overall definitional approach: In the context of analyzing advanced AI governance, there are both advantages and drawbacks to working with form-centric terms. First, we review five potential benefits.

Benefit (1): Well-established and recognized terms: In the first place, using form-centric terms has the advantage that many of these terms are relatively well established and familiar.[73] Out of all the terms surveyed in this report, many form-centric definitions for advanced AI, like strong AI, superintelligence, or AGI, have both the longest track record and the greatest visibility in academic and public debates around advanced AI. Moreover, while some of these terms are relatively niche to philosophical ("AI+") or technical subcommunities ("CAIS"), many of these terms are in fact the ones used prominently by the main labs developing the most disruptive, cutting-edge AI systems.[74] Prima facie, reusing these terms could avoid the problem of having to reinvent the wheel and achieve widespread awareness of and buy-in on newer, more niche terms.

Benefit (2): Readily intuitive concepts: Secondly, form-centric terms evoke certain properties—such as autonomy, adaptability, and human-likeness—which, while certainly not uncontested, may be concepts that are more readily understood or intuited by the public or policymakers than would be more scientifically niche concepts. At the same time, this may also be a drawback, if the ambiguity of many of these terms opens up greater scope for misunderstanding or flawed assumptions to creep into governance debates.

Benefit (3): Enables more forward-looking and anticipatory policymaking towards advanced AI systems and their impacts. Thirdly, because some (though not all) form-centric definitions of advanced AI relate to systems that are perceived (or argued) to appear in the future, using these terms could help extend public attention, debate, and scrutiny to the future impacts of yet more general AI systems which, while their arrival might be uncertain, would likely be enormously impactful. This could help such debates and policies to be less

[73] Though that critically does not mean uncontested or uncontroversial.

[74] 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.

reactive to the impacts of each latest AI model release or incident and start laying the foundations for major policy initiatives. Indeed, centering governance analysis on form-centric terms, even if they are (seen as) futuristic or speculative, can help inform more forward-looking, anticipatory, and participatory policymaking towards the kind of AI systems (and the kind of capabilities and impacts) that may be on the horizon.[75]

One caveat here is that to consider this a benefit, one has to strongly assume that these futuristic forms of advanced AI systems are in fact feasible and likely near in development. At the same time, this approach need not presume absolute certainty over which of these forms of advanced AI can or will be developed, or on what timelines; rather, well-established risk management approaches[76] can warrant some engagement with these scenarios even under uncertainty. To be clear, this need not (and should not) mean neglecting or diminishing policy attention for the impacts of existing AI systems,[77] especially as these impacts are already severe and may continue to scale up as AI systems both become more widely implemented and create hazards for existing communities.

Benefit (4): Enables public debate and scrutiny of overarching (professed) direction and destination for AI development. Fourthly, and relatedly, this above advantage to using form-centric terms could still hold, even if one is very skeptical of these types of futuristic AI, because they afford the democratic value of allowing the public and policymakers to chime in on the actual professed long-term goals and aspirations of many (though not all) leading AI labs.[78]

In this way, the cautious, clear, and reflexive use of terms such as AGI in policy debates could be useful *even if* one is very skeptical of the actual feasibility of these forms of AI (or believes they are possible but remains skeptical that they will be built anytime soon using extant approaches). This is because there is democratic and

[75] 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

[76] See for instance: 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.

[77] Ibid. See also 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. For another argument that models the causes and consequences of disunity within actors focused on different AI issues, see also 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.

[78] Ibid. See also previously 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. For particular cases of such labs (or their researchers) drawing on these terms, see for instance: (for OpenAI) Altman, Sam, Greg Brockman, and Ilya Sutskever. 'Governance of Superintelligence'. OpenAI, 22 May 2023. https://openai.com/blog/governance-of-superintelligence.; (for Microsoft) 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.

procedural value in the public and policymakers being able to hold labs to account for the goals that they in fact espouse and pursue—even if those labs may turn out mistaken about the ability to execute on those plans (in the near term).[79] This is especially the case when these are goals that the public might not (currently) agree with or condone.[80]

Using these "futuristic" terms could therefore help ground public debate over whether the development of these particular systems is even a societal goal they condone, whether society might prefer for labs or society to pursue a different visions for society's relation to AI technology,[81] or (if these systems are indeed considered desirable and legitimate goals) what additional policies or guarantees the world should demand.[82]

Benefit (5): Technology neutrality: Fifthly, the use of form-centric terms in debates can build in a degree of technology neutrality[83] in policy responses, since debates need not focus on the specific engineering or scientific pathways by which one or another highly capable and impactful AI system is pursued or developed. This could make the resulting regulatory frameworks more scalable and future-proof.

At the same time, there are a range of general drawbacks to using (any of these) form-focused definitions in advanced AI governance.

Drawback (1): Connotations and baggage around terms: In the first place, the greater familiarity of some of these terms means that many form-focused terms have become loaded with cultural baggage, associations, or connotations which may mislead, derail, or unduly politicize effective policymaking processes. In particular, many of these terms are contested and have become associated (whether or not necessarily) with particular

[79] See also: Futerman, Gideon. 'We Are Fighting a Shared Battle (a Call for a Different Approach to AI Strategy)'. EA Forum, 16 March 2023. https://forum.effectivealtruism.org/posts/Q4rg6vwbtPxXW6ECj/we-are-fighting-a-shared-battle-a-call-for-a-different.

[80] See for instance: AI Policy Institute. 'Poll Shows Overwhelming Concern About Risks From AI as New Institute Launches to Understand Public Opinion and Advocate for Responsible AI Policies', 9 August 2023. https://theaipi.org/poll-shows-overwhelming-concern-about-risks-from-ai-as-new-institute-launches-to-understand-public-opinion-and-advocate-for-responsible-ai-policies/.

[81] See for instance Matteucci, Kayla, Shahar Avin, Fazl Barez, and Seán Ó hÉigeartaigh. 'AI Systems of Concern'. arXiv, 9 October 2023. https://doi.org/10.48550/arXiv.2310.05876. (discussing alternate paradigms for positive AI futures that focus not on systems such as AGI but rather visions of collective intelligence, human-centered AI, or comprehensive AI services). Referring to 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.

[82] See for instance Axiotes, Connor, and Eddie Bolland. 'Tipping Point: On The Edge of Superintelligence'. Adam Smith Institute, 27 September 2023. https://www.adamsmith.org/research/tipping-point-on-the-edge-of-superintelligence-1. (discussing, amongst others, various policies to implement in response to the labor effects of AGI).

[83] For a discussion of technology neutrality, see also 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 Pg. 408. ("Tech-neutral rules are framed broadly, often with the aim of applying to activities or their consequences regardless of the technology employed [...] One of the main appeals of tech neutrality lies in the intuition that it is more flexible and "future -proof" than those regulating specific technologies."). For a more detailed discussion of different forms of technology-neutral rules, see also: Koops, Bert-Jaap. 'Should ICT Regulation Be Technology-Neutral?' In *Starting Points for ICT Regulation: Deconstructing Prevalent Policy One-Liners*, edited by Bert-Jaap Koops, Miriam Lips, Corien Prins, and Maurice Schellekens, 9:77–108. IT & Law Series. T.M.C. Asser Press, 2006. https://papers.ssrn.com/abstract=918746. (distinguishing four legislative purposes for creating a tech-neutral rule, including achieving specific effects or ensuring that different modes of a particular activity are treated functionally equivalent, minimizing discrimination between technologies, or future-proofing the law).

views or agendas towards building these systems.[84] This is a problem because, as discussed previously, the use of different metaphors, frames, and analogies may be irreducible in (and potentially even essential to) the ways that the public and policymakers make sense of regulatory responses. Yet different analogies—and especially the unreflexive use of terms—also have limits and drawbacks and create risks of inappropriate regulatory responses.[85]

Drawback (2): Significant variance in prominence of terms and constant turnover: In the second place, while some of these terms have held currency at different times in the last decades, many do not see equally common use or recognition in modern debates. For instance, terms such as "strong AI" which dominated early philosophical debates, appear to have fallen slightly out of favor in recent years[86] as the emergence and impact of foundation models generally, and generative AI systems specifically, has revived significantly greater attention to terms such as "AGI". This churn or turnover in definitions may mean that it may not be wise to attempt to pin down a single term or definition right now, since analyses that focus on one particular anticipated form of advanced AI may be more likely to be rendered obsolete. At the same time, this is likely to be a general problem with any concepts or terminology chosen.

Drawback (3): Contested terms, seen as speculative or futuristic: In the third place, while some form-centric terms (such as "GPAIS" or "foundation model") have been well established in AI policy debates or processes, others, such as "AGI", "strong AI", or "superintelligence", are more future-oriented, referring to advanced AI systems that do not (yet) exist.[87] Consequently, many of these terms are contested and seen as futuristic and speculative. This perception may be a challenge, because even if it is incorrect (e.g., such that particular systems like "AGI" will in fact be developed within short timelines or are even in some sense "already here"[88]), the mere perception that a technology or term is far-off or "speculative" can serve to inhibit and delay effective regulatory or policy action.[89]

A related but converse risk of using future-oriented terms for advanced AI policy is that it may inadvertently import a degree of technological determinism[90] in public and policy discussions, as it could imply that one or

[84] Jones, Elliot. 'Explainer: What Is a Foundation Model?' Ada Lovelace Institute, 17 July 2023. https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/. ("Some other terms, such as 'frontier models' and 'AGI/strong AI' are also being used in industry, policy and elsewhere, but are more contested. This is in part because of the lack of a specific interpretation, and in part because of their origins and the context in which they are used").

[85] See also the previous discussions on the risks of analogies.

[86] Although notably this may not be the case across cultures: for instance, the term "strong AI" may potentially be more recognized and used amongst some Chinese AI researchers: see 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.

[87] Although see: Arcas, Blaise Agüera y, and Peter Norvig. 'Artificial General Intelligence Is Already Here'. *Noema*, 10 October 2023. https://www.noemamag.com/artificial-general-intelligence-is-already-here. ("today's frontier models perform competently even on novel tasks they were not trained for, crossing a threshold that previous generations of AI and supervised deep learning systems never managed. Decades from now, they will be recognized as the first true examples of AGI, just as the 1945 ENIAC is now recognized as the first true general-purpose electronic computer").

[88] ibid.

[89] For instance, 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. (discussing the failure of early, mid-2000s efforts to put "killer robots" on the international humanitarian disarmament issue agenda because they were seen as "too speculative").

[90] Notably, there is widespread confusion over the term "technological determinism". For a taxonomy of different uses, see: Dafoe, Allan. 'On Technological Determinism: A Typology, Scope Conditions, and a Mechanism'. *Science, Technology, & Human Values* 40, no. 6 (1 November 2015): 1047–76. https://doi.org/10.1177/0162243915579283. See also Wyatt, Sally. 'Technological Determinism Is Dead; Long Live Technological Determinism'. In *The Handbook of Science and*

another particular forms or architectures of advanced AI ("AGI", "strong AI") are not just possible but inevitable—thereby shifting public and policy discussions away from the question of *whether* we should (or can safely) develop these systems (rather than other, more beneficial architectures)[91] towards less ambitious questions over *how* we should best (safely) reckon with the arrival or development of these technologies.

In response, this drawback could be somewhat mitigated by relying on terms for the forms of advanced AI—such as GPAIS or highly-capable foundation models—that are (a) more present-focused, while (b) not putting any strong presumed ceilings on the capabilities of the systems.

Drawback (4): Definitional ambiguity: In the fourth place, many of these terms, and especially future-oriented terms such as "strong AI", "AGI", and "human-level AI", suffer from definitional ambiguity in that they are used both inconsistently and interchangeably with one another.[92]

Of course, just because there is no settled or uncontested definition for a term such as "AGI" does not make it prima facie unsuitable for policy or public debate. By analogy, the fact that there can be definitional ambiguity over the content or boundaries of concepts such as "the environment" or "energy" does not render "environmental policy" or "energy policy" meaningless categories or irrelevant frameworks for regulation.[93] Nor indeed does outstanding definitional debate mean that any given term, such as AGI, is "meaningless."[94]

Nonetheless, the sheer range of contesting definitions for many of these concepts may reflect an underlying degree of disciplinary or philosophical confusion, or at least suggest that, barring greater conceptual clarification and operationalization,[95] these terms will lead to continued disagreement. Accordingly, anchoring advanced AI governance to broad terms such as "AGI" may make it harder to articulate appropriately scoped legal obligations for specific actors that will not end up being over- or underinclusive.[96]

Technology Studies, edited by Edward J. Hackett, Olga Amsterdamska, Judy Wajcman, Michael Lynch, Anthony Giddens, and Judy Wajcman, 165–80. MIT Press, 2008. See also Peters, John Durham. '"You Mean My Whole Fallacy Is Wrong": On Technological Determinism'. *Representations* 140, no. 1 (1 November 2017): 10–26. https://doi.org/10.1525/rep.2017.140.1.10. (offering a history of the concept of "technological determinism", a discussion of how it has come to be perceived as a "fallacy," and a critique of "the ways academic accusations of fallaciousness risk stopping difficult but essential kinds of inquiry.")

[91] Matteucci, Kayla, Shahar Avin, Fazl Barez, and Seán Ó hÉigeartaigh. 'AI Systems of Concern'. arXiv, 9 October 2023. https://doi.org/10.48550/arXiv.2310.05876.; 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.

[92] See also Appendix 1A.

[93] 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. Pg. 546 ("this challenge is not unique to AI: definitional issues abound in areas such as environment and energy, but have not figured prominently in debates over centralisation. Indeed, energy and environment ministries are common at the domestic level.")

[94] See informally Ricon, Jose Luis. 'Set Sail For Fail? On AI Risk'. *Nintil*, 4 August 2022. https://nintil.com/ai-safety. Appendix A. (discussing and responding to a series of critiques of the concept of AGI and of various components of the AI risk argument).

[95] For one such recent attempt, see: Morris, Meredith Ringel, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 'Levels of AGI: Operationalizing Progress on the Path to AGI'. arXiv, 4 November 2023. https://doi.org/10.48550/arXiv.2311.02462. (articulating a framework for "levels of AGI" that distinguishes AI systems on the basis of both their performance—i.e., whether the system's performance is "emerging, competent, expert, virtuoso, and superhuman"—and their generality—i.e., whether the system is narrow or general).

[96] In the same way that relying on the broad umbrella term "AI" can end up in laws that are over- or underinclusive: Schuett, Jonas. 'Defining the Scope of AI Regulations'. *Law, Innovation and Technology* 15, no. 1 (3 March 2023): 1–23. https://doi.org/10.1080/17579961.2023.2184135.

Drawback (5): Challenges in measurement and evaluation: In the fifth place, an underlying and related challenge for the form-centric approach is that (in part due to these definitional disagreements and in part due to deeper reasons) it faces challenges around how to measure or operationalize (progress towards) advanced AI systems.

This matters because effective regulation or governance—especially at the international level[97]—often requires (scientific and political) consensus around key empirical questions, such as when and how we can know that a certain AI system truly achieves some of the core features (e.g., autonomy, agency, generality, and human-likeness) that are crucial to a given term or concept. In practice, AI researchers often attempt to measure such traits by evaluating an AI system's ability to pass one or more specific benchmark tests (e.g., the Turing test, the Employment test, the SAT, etc.).[98]

However, such testing approaches have many flaws or challenges.[99] At the practical level, there have been problems with how tests are applied and scored[100] and how their results are reported.[101] Underlying this is a challenge that the way in which some common AI performance tests are constructed may emphasize nonlinear or discontinuous metrics, which can lead to an overtly strong impression that some model skills are "suddenly" emergent properties (rather than smoothly improving capabilities).[102] More fundamentally, there have been challenges to the meaningfulness of applying human-centric tests (such as the bar exam) to AI systems[103] and indeed deeper critiques of the construct validity of leading benchmark tests in terms of whether they actually are indicative of progress towards flexible and generalizable AI systems.[104]

[97] Maas, Matthijs M., and José Jaime Villalobos. 'International AI Institutions: A Literature Review of Models, Examples, and Proposals'. *AI Foundations Report 1*. Institute for Law & AI, September 2023. https://www.law-ai.org/international-ai-institutions pg. 13-20 (discussing the role of scientific consensus-building institutions such as the IPCC or political consensus-building institutions such as the G7 or G20 as they are invoked as models for global AI governance).

[98] At least those that focus on an empirical assessment rather than a prescriptive account of how (by what pathways) AI is to be constructed. See for instance Arcas, Blaise Agüera y, and Peter Norvig. 'Artificial General Intelligence Is Already Here'. *Noema*, 10 October 2023. ("For each criticism, we should ask whether it is prescriptive or empirical. A prescriptive criticism would argue: 'In order to be considered as AGI, a system not only has to pass this test, it also has to be constructed in this way.' We would push back against prescriptive criticisms on the grounds that the test itself should be sufficient — and if it is not, the test should be amended").

[99] 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.; significantly and problematically, challenges around evaluation occur not just in testing the progress in AI system capabilities but also in designing effective, robust, and reliable evaluation suites for their safety. For an accessible overview, see: Anthropic. 'Challenges in Evaluating AI Systems'. Anthropic, 4 October 2023. https://www.anthropic.com/index/evaluating-ai-systems.

[100] See for instance: Martínez, Eric. 'Re-Evaluating GPT-4's Bar Exam Performance'. SSRN Scholarly Paper. Rochester, NY, 8 May 2023. https://doi.org/10.2139/ssrn.4441311.

[101] Burnell, Ryan, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, et al. 'Rethink Reporting of Evaluation Results in AI'. *Science* 380, no. 6641 (14 April 2023): 136–38. https://doi.org/10.1126/science.adf6369.

[102] Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 'Are Emergent Abilities of Large Language Models a Mirage?' arXiv, 28 April 2023. https://doi.org/10.48550/arXiv.2304.15004.

[103] Hernandez-Orallo, Jose. 'Beyond the Turing Test'. *Journal of Logic, Language and Information* 9, no. 4 (1 October 2000): 447–66. https://doi.org/10.1023/A:1008367325700.; Hernández-Orallo, José. 'Twenty Years Beyond the Turing Test: Moving Beyond the Human Judges Too'. *Minds and Machines* 30, no. 4 (2020): 533–62. https://doi.org/10.1007/s11023-020-09549-0.

[104] Raji, Inioluwa Deborah, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 'AI and the Everything in the Whole Wide World Benchmark'. arXiv, 26 November 2021. https://doi.org/10.48550/arXiv.2111.15366.

Of course, that does not mean that there may not be further scientific progress towards the operationalization of useful tests for understanding when particular forms of advanced AI such as AGI have been achieved.[105] Nor is it to suggest that benchmark and evaluation challenges are unique to form-centric definitions of AI—indeed, they may also challenge many approaches focused on specific capabilities of advanced AIs.[106] However, the extant challenges over the operationalization of useful tests mean that overreliance on these terms could muddle debates and inhibit consensus over whether a particular advanced system is within reach (or already being deployed).

Drawback (6): Overt focus on technical achievement of particular forms may make this approach underinclusive of societal impacts or capabilities: In the sixth place, the focus of future-oriented form-centric approaches on the realization of one or another type of advanced AI system ("AGI", "human-level AI"), might be adequate if the purpose for our definitions is for *technical research.*[107] However, for those whose definitional purpose is to understand AI's societal impacts (*sociotechnical research*) or to appropriately regulate AI (*regulatory*), many form-centric terms may miss the point.

This is because what matters from the perspective of human and societal safety, welfare, and well-being—and from the perspective of law and regulation[108]—is not the achievement of some fully general capacity in any individual system but rather overall sociotechnical impacts or the emergence of key dangerous

[105] See for instance: Zhong, Wanjun, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 'AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models'. arXiv, 13 April 2023. https://doi.org/10.48550/arXiv.2304.06364.

[106] Indeed, they reflect a general difficulty of applying human-centered tests of intelligence or competence to systems that may display types or forms of intelligence that are not necessarily human—a problem that echoes past debates over the nature (and proper measurement of) intelligence in the animal kingdom. See Long, Robert. 'Are We Smart Enough to Know How Smart AIs Are?' *Asterisk*, 2023. https://asteriskmag.com/issues/03/are-we-smart-enough-to-know-how-smart-ais-are. ("The strangeness of LLMs means that they are smart in their own way. They can neither be presumed to be mere next-token predictors, or to neatly map onto human psychology. As de Waal says of chimpanzees, thinking of large language models only in terms of whether they meet or fail to meet human standards of intelligence does not do them justice. Naive anthropomorphism can give us an inflated view of what they can do. It can also lead us to underestimate them by blinding us to complex and inhuman ways they have of being intelligent").

[107] See also above (section "Three purposes for definitions").

[108] See also Maas, Matthijs M. 'Aligning AI Regulation to Sociotechnical Change'. 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, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.22. (pg. 6-7). (arguing that AI systems can create a regulatory rationale under existing theories of regulation, "whenever [that AI system] drives sociotechnical changes (new ways of carrying out old behavior, or new behaviors, relations, or entities) which result in one or more of the following situations:

1. New possible market failures;
2. New risks to human health or safety, or to the environment;
3. New risks to moral interests, rights, or values;
4. New threats to social solidarity;
5. New threats to democratic process; or
6. New threats to the coherence, efficacy or integrity of the existing regulatory ecosystem charged with mitigating the prior direct risks (1–5)").

This taxonomy draws on: Bennett Moses, Lyria. 'Regulating in the Face of Sociotechnical Change'. In *The Oxford Handbook of Law, Regulation, and Technology*, edited by Roger Brownsword, Eloise Scotford, and Karen Yeung, 573–96, 2017. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-49. Pg. 578.

capabilities—even if they derive from systems that are not yet (fully) general[109] or that develop dangerous emergent capabilities that are not human-like.[110] Given all this, there is a risk that taking a solely form-centric approach leaves advanced AI governance vulnerable to a version of the "AI effect," whereby "real AGI" is always conceived of as being around the corner but rarely as a system already in production.

Suitability of different terms within approach: Given the above, if one does aim to draw on this approach, it may be worth considering which terms manage to gain from the strengths of this approach while reducing some of the pitfalls. In this view, the terms "GPAIS" or "foundation model" may be more suitable in many contexts, as they are recognized as categories of (increasingly) general and competent AI systems of which some versions already exist today. In particular, because (versions) of these terms are already used in ongoing policy debates, they could provide better regulatory handles for governing the development of advanced AI—for instance by their relation to the complex supply chain of modern AI development that contains both upstream and downstream developers and users.[111] Moreover, these terms do not presume a ceiling in the system's capability; accordingly, concepts such as "highly-capable foundation model",[112] "extremely capable foundation model", or "threshold foundation model" could help policy debates be cognizant of the growing capabilities of these systems while still being more easily understandable for policymakers.[113]

2. Development-based definitions: Pathways towards advanced AI

Focus of approach: A second cluster of terms focuses on the anticipated or hypothesized scientific pathways or paradigms that could be used to create advanced AI systems. Notably, the goal or target of these pathways is often to build "AGI"-like systems.[114]

Notes and caveats: Any discussion of proposed pathways towards advanced AI has a number of important

[109] Gruetzemacher, Ross, and Jess Whittlestone. 'The Transformative Potential of Artificial Intelligence'. *Futures* 135 (2022): 102884. https://doi.org/10.1016/j.futures.2021.102884. See also Carlsmith, Joseph. 'Is Power-Seeking AI an Existential Risk?' arXiv, April 2021. http://arxiv.org/abs/2206.13353.

[110] See also the overview of potentially dangerous capabilities 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. As well as in section II.4 on "critical capabilities"

[111] 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/.; See also 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. And generally see: 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/.

[112] 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.

[113] For a similar point, see also Guest, Oliver. 'What Term to Use for AI in Different Policy Contexts?' Effective Altruism Forum, 6 September 2023. https://forum.effectivealtruism.org/posts/9Y5YzNDMdYYg6hjwD/what-term-to-use-for-ai-in-different-policy-contexts.

[114] Another detailed review is given in: 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.

caveats. In the first place, many of these proposed paradigms have long been controversial, with pervasive and ongoing disagreement about their scientific foundations and feasibility as paths towards advanced AI (or in particular as paths towards particular forms of advanced AI, such as AGI).[115] Secondly, these approaches are not necessarily mutually exclusive, and indeed many labs combine elements from several in their research.[116] Thirdly, because the relative and absolute prominence and popularity of many of these paradigms have fluctuated over time and because there are often, as in any scientific field, significant disciplinary gulfs between paradigms, there is highly unequal treatment of these pathways and terms. As such, whereas some paradigms (such as the scaling, reinforcement-learning, and, to some extent, brain-inspired approaches) are reasonably widely known, many of the other approaches and terms listed (such as "seed AI") may be relatively unknown or even very obscure within the modern mainstream machine learning (ML) community.[117]

Other taxonomies: There have been various other such attempts to create taxonomies of the main theorized pathways that have been proposed to build or implement advanced AI. For instance, Goertzel and Pennachin have defined four different approaches to creating "AGI", which to different degrees draw on lessons from the (human) brain or mind.[118] More recently, Hannas and others have drawn on this framework and extended it to five theoretical pathways towards "general AI".[119]

Further extending such frameworks, one can distinguish between at least 11 proposed pathways towards

[115] Heaven, Will Douglas. 'Artificial General Intelligence: Are We Close, and Does It Even Make Sense to Try?' MIT Technology Review, 15 October 2020. https://www.technologyreview.com/2020/10/15/1010461/artificial-general-intelligence-robots-ai-agi-deepmind-google-openai/.

[116] For instance, scaling-based approaches (Table 4) often involve the scaling of particular other approaches (such as reward-based); likewise, modular cognitive architecture approaches can draw from neuro-inspired approaches, amongst others.

[117] In part, this may be because some of these terms focus explicitly on AGI, which has not been an (explicit) focus of many researchers in the modern ML field to date. I thank Lauro Langosco for valuable observations on this point.

[118] Ben Goertzel and Cassio Pennachin, "Contemporary Approaches to Artificial General Intelligence," in Artificial General Intelligence, Goertzel and Pennachin, eds., Springer, Berlin, 2007, pg 22. ("1. approaches that attempt to model biological brains; 2. approaches explicitly guided by the human mind and brain; 3. approaches inspired by the human mind much more than the brain; 4. approaches that depend little on known science about human intelligence").

[119] 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. 4–5). In their view, these are:

1. "Attempt to *understand intelligence with cues from human behavior* and create machine algorithms that emulate it. This has been the majority viewpoint, associated with traditional ML/deep learning.

2. *Reverse-engineer a human brain* on the assumption that what emerges is intelligence. This 'neuromorphic' or brain-imitative approach derives function from structure and is the province of 'brain-inspired AI' and 'connectomics.'

3. Force the emergence of intelligence, in theory, by *running algorithms fast enough to 'recreate the same amount of cumulative optimization power* that the relevant processes of natural selection instantiated throughout our evolutionary past.'

4. *Expand the definition of intelligence*. As we argue above, there is no reason to view intelligence as uniquely human. Any '*de novo*' AI substantially able to achieve wide goals would qualify.

5. Finally, *use brain-computer interfaces* to position both elements, human and machine, to achieve (or overachieve) human goals. Embedded nanoscale chips and high-throughput cognitive 'offloading' (partial brain emulation) are hypothetical approaches." (emphasis added).

An older schema, reproducing many of these categories, is given in: Brin, David. 'How Might AI Come About? Different Approaches and Their Implications for Life in the Universe'. In *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, 2018. https://www.taylorfrancis.com/chapters/edit/10.1201/9781351251389-9/might-ai-come-david-brin advanced AI (See Table 4).

Terms: Many of these paradigms or proposed pathways towards advanced AI come with their own assorted terms and definitions (see Appendix 1B). These terms include amongst others de novo AGI, prosaic AGI, frontier (AI) model [compute threshold], [AGI] from evolution, [AGI] from powerful reinforcement learning agents, powerful deep learning models, seed AI, neuroAI, brain-like AGI, neuromorphic AGI, whole-brain emulation, brain-computer interface, [advanced AI based on] a sophisticated embodied agent, or hybrid AI (see Table 4).

Definitions: As noted, these terms can be mapped on 11 proposed pathways towards advanced AI, with their own terms for the resulting advanced AI systems.

Table 4: Pathway-focused definitions of advanced AI

Pathway	**Broad description**[120]	**Terms** [# of definitions surveyed][121]
First-principles approaches	Based on new fundamental insights in computer science, mathematics, and algorithms, producing systems that may but need not mimic human cognition	→ De novo AGI [1]
Scaling approaches	Based on "brute forcing" advanced AI by running leading, existing AI approaches with more computing power and/or training data, as per the "scaling hypothesis"	→ Prosaic AGI [1] → Frontier (AI) model *[compute-threshold definition]* [2]
Evolutionary approaches	Based on algorithms competing to mimic the evolutionary brute search process that produced human intelligence	→ [AGI] from evolution [1]
Reward-based approaches	Based on running reinforcement learning systems with simple rewards in rich environments	→ [AGI] from powerful reinforcement learning agents [1] → Powerful deep learning models [1]
Bootstrapping approaches	Approaches that pursue a minimally intelligent core system capable of subsequent recursive self-improvement or improvement through leveraging hardware or data overhangs	→ Seed AI [3]

[120] For sources and references for each of these approaches, see Appendix 1B.

[121] For the specific definitions and sources, see Appendix 1B.

Neuro-inspired approaches	Various forms of biologically-inspired, brain-inspired, or brain-imitative approaches that draw on neuroscience and "connectomics" to reproduce general intelligence.	→ NeuroAI [1] → Brain-like AGI [1] → Neuromorphic AGI [1]
Neuro-emulated approaches	Digitally simulate or recreate the states of human brains at fine-grained level.	→ Whole-brain emulation [2] → Digital people *[emulation definition]* [1]
Neuro-integrationist approaches	Based on merging components of human and digital cognition	→ Brain-computer interfaces (BCI) [1]
Embodiment approaches	Based on providing the AI system with a robotic physical "body" to ground cognition and enable it to learn from direct experience of the world	→ Embodied agent [1]
Modular cognitive architecture approaches	Used in various fields but especially in robotics, where researchers integrate well-tested but "frozen" state-of-the-art modules (perception, reasoning, etc.) to improve agent performance without independent learning	→ (no clear single term)
Hybrid approaches	Approaches that rely on combining deep neural network-based approaches to AI with other paradigms (such as symbolic AI)	→ Hybrid AI [1]

Notably, there are significant differences in the prominence of these approaches—and the resources dedicated to them—at different frontier AI labs today. For instance, while some early work on the governance of advanced AI systems focused on AI systems that would (presumably) be built from first principles, bootstrapping,[122] or neuro-emulated approaches (see Table 4), much of such work has more recently shifted to focus on understanding the risks from and pathways to aligning and governing advanced AI systems created through computational scaling.

This follows high-profile trends in leading AI labs. While (as discussed above) many research labs are not dedicated to a single paradigm, the last few years (and 2023 in particular) have seen a significant share of resources going towards computational scaling approaches, which have yielded remarkably robust (though not uncontested) performance improvements.[123] As a result, the scaling approach has been prominent in informing

[122] See for instance Yudkowsky, Eliezer. 'General Intelligence and Seed AI'. Singularity Institute, 2001. https://web.archive.org/web/20120805130100/singularity.org/files/GISAI.html.

[123] Bowman, Samuel R. 'Eight Things to Know about Large Language Models', 2023. https://cims.nyu.edu/~sbowman/eightthings.pdf.

the approaches of labs such as OpenAI,[124] Anthropic,[125] DeepMind,[126] and Google Brain (now merged into Google DeepMind).[127] This approach has also been prominent (though somewhat lagging) in some Chinese labs such as Baidu, Alibaba, Tencent, and the Beijing Institute for General Artificial Intelligence.[128] Nonetheless, other approaches continue to be in use. For instance, neuro-inspired approaches have been prominent in DeepMind,[129] Meta AI Research,[130] and some Chinese[131] and Japanese labs,[132] and modular cognitive architecture approaches have informed the work by Goertzel's OpenCog project,[133] amongst others.

Suitability of overall definitional approach: In the context of analyzing advanced AI governance, there are both advantages and drawbacks to using concepts that focus on pathways of development.

Amongst the advantages of this approach are:

Benefit (1): Close(r) grounding in actual technical research agendas aimed at advanced AI: Defining advanced AI systems according to their (envisioned) development pathways has the benefit of keeping advanced AI governance debates more closely grounded in existing technical research agendas and programs, rather than the often more philosophical or ambiguous debates over the expected *forms* of advanced AI systems.

[124] Branwen, Gwern. 'The Scaling Hypothesis', 28 May 2020. https://www.gwern.net/Scaling-hypothesis. And for a bibliographic overview, see: Branwen, Gwern. 'Machine Learning Scaling', 24 April 2021. https://www.gwern.net/notes/Scaling.

[125] Anthropic. 'Research Principles'. Accessed 29 November 2022. https://www.anthropic.com/#research-principles.

[126] Branwen, Gwern. 'The Scaling Hypothesis', 28 May 2020. https://www.gwern.net/Scaling-hypothesis. ("DeepMind holds what we might call the 'weak scaling hypothesis': they believe that AGI will require us to 'find the right algorithms' effectively replicating a mammalian brain module by module, and that while these modules will be extremely large & expensive by contemporary standards (which is why compute is important, to give us 'a more powerful tool with which to hunt for the right algorithms'), they still need to be invented & fine tuned piece by piece, with little risk or surprise until the final assembly.").

[127] Pichai, Sundar. 'Google DeepMind: Bringing Together Two World-Class AI Teams'. Google, 20 April 2023. https://blog.google/technology/ai/april-ai-update/.

[128] 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. 7-11.

[129] Branwen, Gwern. 'The Scaling Hypothesis'.

[130] LeCun, Yann. 'A Path Towards Autonomous Machine Intelligence'. *OpenReview*, 27 June 2022. https://openreview.net/forum?id=BZ5a1r-kVsf.

[131] 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/.; See also 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 12–16).

[132] Ryota Kanai, for instance, and 'The Whole Brain Architecture Initiative', 14 September 2015. https://wba-initiative.org/en/. I thank José Hernández-Orallo for this suggestion.

[133] 'OpenCog'. Accessed 24 May 2023. https://wiki.opencog.org/w/The_Open_Cognition_Project.; as discussed in: Heaven, Will Douglas. 'Artificial General Intelligence: Are We Close, and Does It Even Make Sense to Try?' MIT Technology Review, 15 October 2020. https://www.technologyreview.com/2020/10/15/1010461/artificial-general-intelligence-robots-ai-agi-deepmind-google-openai/.

Benefit (2): Technological specificity allowing scoping of regulation to approaches of concern: Relatedly, this also allows better regulatory scoping of the systems of concern. After all, the past decade has seen a huge variety amongst AI techniques and approaches, not just in terms of their efficacy but also in terms of the issues they raise, with particular technical approaches raising distinct (safety, interpretability, robustness) issues.[134] At the same time, these correlations might be less relevant in the last few years given the success of scaling-based approaches at creating remarkably versatile and general-purpose systems.

However, taking the pathways-focused approach to defining advanced AI has its own challenges:

Drawback (1): Brittleness as technological specificity imports assumptions about pathways towards advanced AI: The pathway-centric approach may import strong assumptions about what the relevant pathways towards advanced AI are. As such, governance on this basis may not be robust to ongoing changes or shifts in the field.

Drawback (2): Suitability of terms within this approach: Given this, development-based definitions of pathways towards advanced AI seem particularly valuable if the purpose of definition is *technical research* but may be less relevant if the purpose is *sociotechnical analysis* or *regulation.* Technical definitions of AI might therefore provide an important baseline or touchstone for analysis in many other disciplines, but they may not be fully sufficient or analytically enlightening to many fields of study dealing with the societal consequences of the technology's application or with avenues for governing these.

At any rate, one interesting feature of development-based definitions of advanced AI is that the choice of approach (and term) to focus on has significant and obvious downstream implications for framing the policy agendas for advanced AI—in terms of the policy issues to address, the regulatory "surface" of advanced AI (e.g., the necessary inputs or resources to pursue research along a certain pathway), and the most feasible or appropriate tools. For instance, a focus on neuro-integrationist-produced brain-computer interfaces suggests that policy issues for advanced AI will focus less on questions of value alignment[135] and rather around (biomedical) questions of human consent, liability, privacy, (employer) neurosurveillance,[136] and/or morphological freedom.[137] A focus on embodiment-based approaches towards robotic agents raises more established debates from robot law.[138] Conversely, if one expects that the pathway towards advanced AI still requires underlying scientific breakthroughs, either from first principles or through a hybrid approach, this would imply that very powerful AI systems could be developed suddenly by small teams or labs, which lack large compute budgets.

[134] See e.g. Hernandez-Orallo, Jose, Fernando Martınez-Plumed, Shahar Avin, Jess Whittlestone, and Seán Ó hÉigeartaigh. 'AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues'. In *European Conference on Artificial Intelligence*, 8, 2020. https://ecai2020.eu/papers/1364_paper.pdf.

[135] However, see also: Rafferty, Jack. 'Brain-Computer Interfaces: A New Existential Risk Factor'. *Journal of Futures Studies*, 2021. http://jfsdigital.org/brain-computer-interfaces-a-new-existential-risk-factor/.

[136] Muhl, Ekaterina, and Roberto Andorno. 'Neurosurveillance in the Workplace: Do Employers Have the Right to Monitor Employees' Minds?' *Frontiers in Human Dynamics* 5 (2023). https://www.frontiersin.org/articles/10.3389/fhumd.2023.1245619.; Maiseli, Baraka, Abdi T. Abdalla, Libe V. Massawe, Mercy Mbise, Khadija Mkocha, Nassor Ally Nassor, Moses Ismail, James Michael, and Samwel Kimambo. 'Brain–Computer Interface: Trend, Challenges, and Threats'. *Brain Informatics* 10, no. 1 (4 August 2023): 20. https://doi.org/10.1186/s40708-023-00199-3.

[137] Sandberg, Anders. 'Morphological Freedom: What Are the Limits to Transforming the Body?', 2017. http://aleph.se/papers/MF2.pdf.

[138] Calo, Ryan. 'Robotics and the Lessons of Cyberlaw'. *California Law Review* 103 (2015): 513–64. https://heinonline.org/HOL/P?h=hein.journals/calr103&i=539

Similarly, focusing on scaling-based approaches—which seems most suitable given the prominence and success of this approach in driving the recent wave of AI progress—leads to a "compute-based" perspective on the impacts of advanced AI.[139] This suggests that the key tools and levers for effective governance should focus on compute governance—provided we assume that this will remain a relevant or feasible precondition for developing frontier AI. For instance, such an approach underpins the compute-threshold definition for frontier AI, which defines advanced AI with reference to particular technical elements or inputs (such as a compute usage or FLOP threshold, dataset size, or parameter count) used in its development.[140] While a useful referent, this may be an unstable proxy given that it may not reliably or stably correspond to the particular capabilities of concern.

3. Sociotechnical-change based definitions: Societal impacts of advanced AI

Focus of approach: A third cluster of definitions in advanced AI governance mostly brackets out philosophical questions of the precise form of AI systems or engineering questions of the scientific pathways towards their development. Rather, it aims at defining advanced AI in terms of different levels of societal impacts.

Many concepts in this approach have emerged from scholarship that aimed to abstract away from these architectural questions and rather explore the aggregate societal impacts of advanced AI. This includes work on AI technology's international, geopolitical impacts[141] as well as work on identifying relevant historical precedents for the technology's societal impacts, strategic stakes, and political economy.[142] Examples of this work are those that identified novel categories of unintended "structural" risks from AI as distinct from "misuse" or "accident" risks,[143] or taxonomies of the different "problem logics" created by AI systems.[144]

Terms: The societal-impact-centric approach to defining advanced AI includes a variety of terms, including: (strategic) general-purpose technology, general-purpose military transformation, transformative AI, radically transformative AI, AGI (economic competitiveness definition), and machine superintelligence.

[139] Barnett, Matthew. 'A Compute-Based Framework for Thinking about the Future of AI', 1 June 2023. https://forum.effectivealtruism.org/posts/fsaogRokXxby6LFd7/a-compute-based-framework-for-thinking-about-the-future-of.

[140] For compute-threshold-based definitions of frontier AI, see also Appendix 1B.

[141] 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/.

[142] 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.; 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.

[143] 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.

[144] Maas, Matthijs M. 'Aligning AI Regulation to Sociotechnical Change'. 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, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.22. (distinguishing between (1) ethical challenges, (2) security threats, (3) safety risks, (4) structural shifts, (5) public good opportunities, and (6) disruption of governance).

Definitions and themes: While many of these terms are subject to a wide range of different definitions (see Appendix 1C), they again feature a range of common themes or patterns (see Table 5).

Table 5: Societal-impact-focused definitions of advanced AI

Term [# of definitions surveyed][145]	**Selected themes and patterns**
(Strategic) general-purpose technology (GPT) [2]	→ Potential of AI to deliver significant economic value and affect national security; therefore of central political interest → AI's societal impact is enormous but not historically unprecedented; impact comparable with past technological revolutions, such as electricity, internal combustion engine, and computers
General-purpose military transformation (GMT) [1]	→ Significant effect on military innovations and industrial productivity growth → Protracted, gradual process
Transformative AI (TAI) [6]	→ Significant, irreversible changes broad enough to impact all of society; possibly precipitates a qualitatively different future → Transition comparable with the agricultural or industrial revolutions
Radically transformative AI (RTAI) [1]	→ Leads to radical changes to the metrics used to measure human progress and well-being; potential reversal of societal trends previously thought irreversible
Artificial general intelligence (AGI) *[economic competitiveness definition]* [3]	→ Systems can outperform humans at most economically valuable work
Machine superintelligence [1] *[form & impact definition]*	→ Exceeds cognitive capacities of humans → Brings about revolutionary technological and economic advances on very short timescales

Suitability of approach: Concepts within the sociotechnical-change-based approach may be unsuitable if the definitional purpose is *technical research* but particularly appropriate if the purpose is *sociotechnical research*. However, there are a range of complicated benefits and drawbacks to this approach.

Benefit (1): Focus on overall societal impacts rather than technical breakthroughs: One benefit is that sociotechnical-change-based concepts are more appropriate for understanding or debating the broad over-time

[145] For the specific definitions, see Appendix 1C.

societal impacts of advanced AI[146] and how these bend the larger trajectory of society[147] rather than getting bogged down in debates over which of several specific technical pathways will most likely yield such systems or discussions over whether or when the world will achieve a single system with a particular form.[148]

Benefit (2): Grounds impacts of AI in historical precedents: Another potential benefit of these terms is that many of them compare advanced AI systems to established categories of technology (e.g., "general-purpose technologies") or to their impacts (e.g., "industrial revolution"), thereby allowing a clearer discussion of potentially relevant historical precedents for advanced AI systems and a more empirically informed understanding of what might be their societal impacts as well as the resulting political stakes and conditions for governing this technology.[149]

Drawback (1): Reactive and hard to operationalize in advance: The fact that many of these terms bracket the technical features, form, or capabilities of AI systems also creates a risk, however: it may mean that the designation of systems as having a certain impact (i.e., to be "transformative") may be something that can only be effectively assessed in retrospect once a particular set of AI models have had the time to become widely deployed throughout the economy. This may make such definitions somewhat reactive and less suitable for *regulatory* purposes.

Drawback (2): More obscure and jargon-heavy: Another downside or potential risk of this approach is that many of the terms are less well-known outside academic debates.

Drawback (3): Close historical analogies may also mislead or misframe: While there are key benefits to utilizing definitions for AI that may support a learning from historical analogies, the drawback is that these analogies may break down under some assumptions or at least fail to fully capture the impacts (or even simply the features) of advanced AI systems.[150]

[146] See also: Smuha, Nathalie A. 'Beyond the Individual: Governing AI's Societal Harm'. *Internet Policy Review* 10, no. 3 (30 September 2021). https://policyreview.info/articles/analysis/beyond-individual-governing-ais-societal-harm.; and see Clarke, Sam, Jess Whittlestone, Matthijs Maas, Haydn Belfield, Jose Hernandez-Orallo, and Seán Ó HÉigeartaigh. 'Submission of Feedback to the European Commission's Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence'. University of Cambridge (Leverhulme Centre for the Future of Intelligence and Centre for the Study of Existential Risk), 6 August 2021. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665626_en.

[147] For a discussion of such impacts, see: 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. For an older taxonomy of risks, see: Turchin, Alexey, and David Denkenberger. 'Classification of Global Catastrophic Risks Connected with Artificial Intelligence'. *AI and Society* 35, no. 1 (2020): 147–63. https://doi.org/10.1007/s00146-018-0845-5.

[148] See also Barnett, Matthew. 'A Compute-Based Framework for Thinking about the Future of AI', 1 June 2023. https://forum.effectivealtruism.org/posts/fsaogRokXxby6LFd7/a-compute-based-framework-for-thinking-about-the-future-of. ("it's not immediately obvious why we should care about the arrival of a single software system with certain properties. Plausibly, a set of narrow software programs could drastically change the world before the arrival of any monolithic AGI system [...]. In general, it seems more useful to characterize AI timelines in terms of the impacts AI will have on the world.").

[149] Gruetzemacher, Ross, and Jess Whittlestone. 'The Transformative Potential of Artificial Intelligence'. *Futures* 135 (2022): 102884. https://doi.org/10.1016/j.futures.2021.102884.

[150] See e.g. 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/.

4. Risk-based definitions: Critical capabilities achieved by advanced AIs

Focus of approach: Finally, a fourth cluster of terms follows a risk-based approach and focuses on critical capabilities, which certain types of advanced AI systems (whatever their underlying form or scientific architecture) might achieve or enable for human users. The development of such capabilities could then mark key thresholds or inflection points in the trajectory of society.

Other taxonomies: Work focused on the significant potential impacts or risks of advanced AI systems is of course hardly new.[151] Yet in the past years, as AI capabilities have progressed, there has been renewed and growing concern that these advances are beginning to create key threshold moments where sophisticated AI systems develop capabilities that allow them to achieve or enable highly disruptive impacts in particular domains, resulting in significant societal risks. These risks may be as diverse as the capabilities in question—and indeed discussions of these risks do not always or even mostly presume (as do many *form*-centric approaches) the development of *general* capabilities in AI.[152] For instance, many argue that existing AI systems may already contribute to catastrophic risks in various domains:[153] large language models (LLMs) and automated biological design tools (BDTs) may already be used to enable weaponization and misuse of biological agents,[154] the military use of AI systems in diverse roles may inadvertently affect strategic stability and contribute to the risk of nuclear escalation,[155] and existing AI systems' use in enabling granular and at-scale monitoring and surveillance[156] may already be sufficient to contribute to the rise of "digital authoritarianism"[157] or "AI-tocracy"[158], to give a few examples.

As AI systems become increasingly advanced, they may steadily and increasingly achieve or enable further critical capabilities in different domains that could be of special significance. Indeed, as leading LLM-based AI

[151] Burden, John, Sam Clarke, and Jess Whittlestone. 'From Turing's Speculations to an Academic Discipline: A History of AI Existential Safety', 23 August 2023, 201–36. https://doi.org/10.11647/obp.0336.09.; For earlier reviews, see: 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.

[152] Indeed, see informally Chapman, David. *Better without AI*, 2023. https://betterwithout.ai/. (coining the term "scary AI" to refer to artificial intelligence systems that are "dramatically more dangerous," and distinguishing this from debates over whether such systems are more human-like or mind-like).

[153] Bucknall, Benjamin S., and Shiri Dori-Hacohen. 'Current and Near-Term AI as a Potential Existential Risk Factor'. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 119–29. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3514094.3534146.

[154] 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.

[155] Maas, Matthijs, Kayla Lucero-Matteucci, and Di Cooke. 'Military Artificial Intelligence as a Contributor to Global Catastrophic Risk'. In *The Era of Global Risk*, 237–84. Open Book Publishers, 2023. https://www.openbookpublishers.com/books/10.11647/obp.0336/chapters/10.11647/obp.0336.10. See also Johnson, James. 'Inadvertent Escalation in the Age of Intelligence Machines: A New Model for Nuclear Risk in the Digital Age'. *European Journal of International Security*, 15 October 2021, 1–23. https://doi.org/10.1017/eis.2021.23.

[156] 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.

[157] 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.

[158] 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.; however, for critiques suggesting that the utility of AI systems in empowering authoritarian states may be overstated or fundamentally limited, see Farrell, Henry, Abraham Newman, and Jeremy Wallace. 'Spirals of Delusion'. *Foreign Affairs*, 2022. https://www.foreignaffairs.com/world/spirals-delusion-artificial-intelligence-decision-making.

systems have advanced in their general-purpose abilities, they have frequently demonstrated emergent abilities that are surprising even to their developers.[159] This has led to growing concern that as these models continue to be scaled up[160] some next generation of these systems could develop unexpected but highly dangerous capabilities if not cautiously evaluated.[161]

What are these critical capabilities?[162] In some existing taxonomies, critical capabilities could include AI systems reaching key levels of performance in domains such as cyber-offense, deception, persuasion and manipulation, political strategy, building or gaining access to weapons, long-horizon planning, building new AI systems, situational awareness, self-proliferation, censorship, or surveillance,[163] amongst others. Other experts have been concerned about cases where AI systems display increasing tendencies and aptitudes towards controlling or power-seeking behavior.[164] Other overviews identify other sets of hazardous capabilities.[165] In all these cases, the concern is that advanced AI systems that achieve these capabilities (regardless of whether they are fully general, autonomous, etc.) could enable catastrophic misuse by human owners, or could demonstrate unexpected extreme—even hazardous—behavior, even against the intentions of their human principals.

Terms: Within the risk-based approach, there are a range of domains that could be upset by critical capabilities. A brief survey (see Table 6) can identify at least eight such capability domains—moral/philosophical,

[159] 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.; Bowman, Samuel R. 'Eight Things to Know about Large Language Models', 2023. https://cims.nyu.edu/~sbowman/eightthings.pdf.

[160] Pistillo, Matteo. 'Compute Governance Key Concepts', (forthcoming draft). An analysis of how such capabilities might scale is also central to Anthropic's recently published framework of "AI Safety Levels" (ASL). Anthropic. 'Anthropic's Responsible Scaling Policy, Version 1.0', 19 September 2023.

[161] 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. See also: Anderljung, Markus, and Paul Scharre. 'How to Prevent an AI Catastrophe'. *Foreign Affairs*, 14 August 2023. https://www.foreignaffairs.com/world/how-prevent-ai-catastrophe-artificial-intelligence.

[162] A different taxonomy that focuses more on reviewing work on the risks from AGI systems can be found in McLean, Scott, Gemma J. M. Read, Jason Thompson, Chris Baber, Neville A. Stanton, and Paul M. Salmon. 'The Risks Associated with Artificial General Intelligence: A Systematic Review'. *Journal of Experimental & Theoretical Artificial Intelligence* 0, no. 0 (13 August 2021): 1–15. https://doi.org/10.1080/0952813X.2021.1964003.

[163] 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. Pg. 6-12. For a survey of recent cases of AI systems demonstrating deceptive behavior, see: 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.; for a discussion of avenues by which LLMs can aid spear phishing hacking attacks, see: Hazell, Julian. 'Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns'. arXiv, 12 May 2023. https://doi.org/10.48550/arXiv.2305.06972.

[164] 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.; Krakovna, Victoria, and Janos Kramar. 'Power-Seeking Can Be Probable and Predictive for Trained Agents'. arXiv, 13 April 2023. https://doi.org/10.48550/arXiv.2304.06528.

[165] See for instance: Hendrycks, Dan, and Mantas Mazeika. 'X-Risk Analysis for AI Research'. arXiv, 21 July 2022. http://arxiv.org/abs/2206.05862. Pg. 13-14 (reviewing 10 hazardous capabilities); 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.

economic, legal, scientific, strategic or military, political, exponential, and (extremely) dangerous.[166] Namely, these include:[167]

- → Concepts relating to AI systems that achieve or enable critical moral and/or philosophical capabilities include artificial/machine consciousness, digital minds, digital people, sentient artificial intelligence, robot rights catastrophe, (negative) synthetic phenomenology, suffering risks, and adversarial technological maturity.
- → Concepts relating to AI systems that achieve or enable critical economic capabilities include high-level machine intelligence, tech company singularity, and artificial capable intelligence (ACI).
- → Concepts relating to AI systems that achieve or enable critical legal capabilities include advanced artificial judicial intelligence, technological-legal lock-in, and legal singularity.
- → Concepts relating to AI systems that achieve or enable critical scientific capabilities include process-automating science and technology and scientist model.
- → Concepts relating to AI systems that achieve or enable critical strategic and/or military capabilities include decisive strategic advantage and singleton.
- → Concepts relating to AI systems that achieve or enable critical political capabilities include stable totalitarianism, value lock-in, and actually existing AI.
- → Concepts related to AI systems that achieve or enable critical exponential capabilities include intelligence explosion, autonomous replication in the real world, autonomous AI research, and duplicator.
- → Concepts relating to AI systems that achieve or enable (extremely) hazardous capabilities include advanced AI, high-risk AI systems, AI systems of concern, prepotent AI, APS systems, WIDGET, rogue AI, runaway AI, frontier (AI) model (under two definitional thresholds), and highly-capable systems of concern.

Definitions and themes: As noted, many of these terms have different definitions (see Appendix 1D). Nonetheless, a range of common themes and patterns can be distilled (see Table 6).

Table 6: Critical capability-focused definitions of advanced AI

Domain	**Terms for critical capabilities** [# of definitions surveyed][168]	**Selected themes and patterns**
Moral and/or philosophical capabilities	→ Artificial/machine consciousness [2] → Digital minds [1] → Digital people *[capability definition]* [1] → Sentient artificial intelligence [1] → Robot rights catastrophe [1] → (Negative) synthetic phenomenology [1] → Suffering risks [1]	→ Systems that achieve morally relevant properties → Resulting risks of a moral crisis or catastrophe, either by default or if protections are under- or overextended

[166] Within these terms, there is some variance (or often, slippage) between the actual referents of these terms, with some connoting (a) the specific capabilities, others referring to (b) the particular types of advanced AI systems that would possess these capabilities, and yet others to (c) the eventual risk outcome for society if these critical capabilities are deployed to that specific domain.

[167] For literature and sources for each of these terms, see Appendix 1D.

[168] For the specific definitions, and sources, see Appendix 1D.

	→ Adversarial technological maturity [2]	
Economic capabilities	→ High-level machine intelligence (HLMI) [3] → Tech company singularity [1] → Artificial capable intelligence (ACI) [2]	→ Systems that achieve capabilities sufficient to become economically competitive at diverse tasks, either at the level of individual workers or at the level of companies → Risks of massive economic displacement
Legal capabilities	→ Advanced artificial judicial intelligence (AAJI) [1] → Technological-legal lock-in [1] → Legal singularity [1]	→ Systems that achieve key performance in legal or judicial decision-making → Risks of legal stagnation and loss of judicial legitimacy
Scientific capabilities	→ Process-automating science and technology (PASTA) [1] → Scientist model [1]	→ Systems that achieve key scientific research capabilities → Risks of sped-up development of potentially hazardous technologies
Strategic or military capabilities	→ "Decisive strategic advantage" [1] → "Singleton" [1]	→ Systems that achieve key strategic decision-making competences → Risks of geopolitical destabilization
Political capabilities	→ Stable totalitarianism [1] → Value lock-in [2] → Actually existing AI (AEAI) [1]	→ Systems that achieve capabilities that enable effective surveillance, propaganda, or other social control → Risks of domination of society by narrow interests
Exponential capabilities	→ Intelligence explosion [2] → Autonomous replication in the real world [1] → Autonomous AI research [1] → Duplicator [1]	→ Systems that achieve key (self)-improvement or -replication capabilities → Risks of sudden amplification of AI systems' activities and impacts

(Extremely) hazardous capabilities[169]	→ “Advanced AI” [2] → “High-risk AI system” [3] → “AI **systems of concern**” [1] → Prepotent AI [1] → APS systems (“power-seeking AI”) [1] → WIDGET [1] → Rogue AI” [2] → Runaway AI [1] → Frontier (AI) model *[relative-capabilities threshold]* [2] → Frontier (AI) model *[dangerous-capabilities threshold]* [2] → Highly-capable systems of concern [1]	→ Systems that achieve key capabilities in various domains → Risks of massive physical and societal-scale harms

Suitability of approach: There are a range of benefits and drawbacks to defining advanced AI systems by their (critical) capabilities. These include (in no particular order):

Benefit (1): Focuses on key capability development points of most concern: A first benefit of adopting the risk-based definitional approach is that these concepts can be used, alone or in combination, to focus on the key thresholds or transition points in AI development that we most care about—not just the aggregate eventual, long-range *societal outcomes* nor the (eventual) “final” *form* of advanced AI, but rather the key intermediate (technical) *capabilities* that would suffice to create (or enable actors to achieve) significant societal impacts: the points of no return.

Benefit (2): Highlighting risks and capabilities can more precisely inform the public understanding: Ensuring that terms for advanced AI systems clearly center on particular risks or capabilities can help the public and policymakers understand the risks or challenges to be avoided, in a way that is far clearer than terms that focus on very general abilities or which are highly technical (i.e., terms within *essence-* or *development*-based approaches, respectively). Such terms may also assist the public in comparing the risks of one model to those posed by another.[170]

Benefit (3): Generally (but not universally) clearer or more concrete: While some terms within this approach are quite vague (e.g., “singleton”) or potentially difficult to operationalize or test for (e.g., “artificial consciousness”), some of the more specific and narrow terms within this approach could offer more clarity, and less definitional drift, to regulation. While many of them would need significant further clarification before they could be suitable for use in legislative texts (whether domestic laws or international treaties), they may offer the basis for more circumscribed, tightly defined professional cornerstone concepts for such regulation.[171]

However, there are also a number of potential drawbacks to risk-based definitions.

[169] That is, systems that pose or enable critical levels of (extreme or existential) risk, regardless of whether they demonstrate a full range of human-level/like cognitive abilities.

[170] I thank Kevin Frazier for insightful observations on this point.

[171] As one analogy, one can consider the case of nuclear energy regulation: while general parlance refers to nuclear power stations as nuclear reactors or nuclear power projects, legislation on nuclear energy regulation and liability often refers to more circumscribed professional terms such as “nuclear installation,” which allows more discrete and targeted policymaking for individual technological artifacts or projects. I thank Aishwarya Saxena for this suggestion.

Drawback (1): Epistemic challenges around "unknown unknown" critical capabilities: One general challenge to this risk-based approach for characterizing advanced AI is that, in the absence of more specific and empirical work, it can be hard to identify and enumerate all relevant risk capabilities in advance (or to know that we have done so). Indeed, aiming to exhaustively list out all key capabilities to watch for could be a futile exercise to undertake.[172] At the same time, this is a challenge that is arguably faced in any domain of (technology) risk mitigation, and it does not mean that doing such analysis to the best of our abilities is void. However, this challenge does create an additional hurdle for regulation, as it heightens the chance that if the risk profile of the technology rapidly changes, regulators or existing legal frameworks will be unsure of how or where to classify that model.

Drawback (2): Challenges around comparing or prioritizing between risk capabilities: A related challenge lies in the difficulty of knowing which (potential) capabilities to prioritize for regulation and policy. However, that need not be a general argument against this approach. Instead, it may simply help us make explicit the normative and ethical debates over what challenges to avoid and prioritize.

Drawback (3): Utilizing many parallel terms focused on different risks can increase confusion: One risk for this approach is that while the use of many different terms for advanced AI systems, depending on their specific critical capabilities in particular domains, can make more for appropriate and context-sensitive discussions (and regulation) within those domains, at an aggregate level this may increase the range of terms that regulators and the public have to reckon with and compare between—with the risk that these actors simply drown in the range of terms.

Drawback (4): Counterproductive connotations of some terms: The risk-capabilities-focused approach contains a range of concepts or terms that may have significant cultural baggage or connotations, which may make them less suitable for policy. This includes terms such as "machine consciousness" or "digital people". Other terms may run the risk that they import connotations or frames that are counter to some of their intended risk-mitigation goals. Take the term "frontier AI model": on the one hand, this term may (to some audiences) signal the danger or uncertainty that might come from entering into new terrain "beyond the known frontier." On the other hand, to many audiences, the framing of AI research in a "frontier" may rather imply operating within a wild, unregulated space, one subject to constant, continued, decentralized expansion rather than considered and deliberated navigation (or even strategic halting at agreed-upon boundaries).

Drawback (5): Outstanding disagreements over appropriate operationalization of capabilities: One further challenge with these terms may lie in the way that some key terms remain contested or debated—and that even clearer terms are not without challenge. For instance, in 2023, the concept of "frontier model" has become subject to increasing debate over its potential adequacy for regulation.[173] Notably, there are at least three ways of operationalizing this concept. The first, *computational threshold*, has been discussed above.[174]

However, a second operationalization for frontier AI focuses on some *relative-capabilities threshold*. This approach includes recent proposals to define "frontier AI models" in terms of capabilities *relative* to other AI

[172] See previously Tasioulas, John. 'First Steps Towards an Ethics of Robots and Artificial Intelligence'. *Journal of Practical Ethics* 7, no. 1 (June 2019): 61–95. http://www.jpe.ox.ac.uk/papers/first-steps-towards-an-ethics-of-robots-and-artificial-intelligence/ pg. 69. (making this argument against many early attempts to draw up and apply lists of ethical principles for AI systems).

[173] Henshall, Will. 'The Heated Debate Over Who Should Control Access to AI'. Time, 25 August 2023. https://time.com/6308604/meta-ai-access-open-source/.

[174] See also section II(2).

systems,[175] as models that "exceed the capabilities currently present in the most advanced existing models" or as "models that are both (a) close to, or exceeding, the average capabilities of the most capable existing models."[176] Taking such a comparative approach to defining advanced AI may be useful in combating the easy tendency of observers to normalize or become used to the rapid pace of AI capability progress.[177] Yet there may be risks with such a comparative approach, especially when tied to a moving wavefront of "the most capable" existing models, as this could easily impose a need on regulators to engage in constant regulatory updating, as well as creating risks of underinclusivity of some foundation models that did not display hazardous capabilities in their initial evaluations, but which once deployed or shared might be reused or recombined in ways that could create or enable significant harms.[178] The risk of embedding this definition of frontier AI in regulation, would be to leave a regulatory gap around significantly harmful capabilities, especially those that are no longer at the technical "frontier," but which remain unaddressed even so. Indeed, for similar reasons, Seger and others have advocated using the concept "highly-capable foundation models" instead.[179]

A third approach to defining frontier AI models has instead focused more on identifying a set of static and absolute criteria grounded in particular dangerous capabilities (i.e., a *dangerous-capabilities threshold*). Such definitions might be useful insofar as they help regulators or consumers identify better when a model crosses a safety threshold and in a way that is less susceptible to slippage or change over time. This could make such concepts more suitable (and resulting regulations less at risk of obsolescence or governance misspecification) than operationalizations of "frontier AI model" that rely on indirect technological metrics (such as compute thresholds) as proxies for these capabilities. Even so, as discussed above, anchoring the "frontier AI model" concept on particular dangerous capabilities leaves open questions around how to best operationalize and create evaluation suites that are able to identify or predict such capabilities ex ante.

Given this, while the risk-based approach may be the most promising ground for defining advanced AI systems from a *regulatory* perspective, it is clear that not all terms in use in this approach are equally suitable, and many require further operationalization and clarification.

[175] Interestingly, in doing so it provides an interesting mirror to existing (*form*-centric) attempts to establish definitions for advanced AI (e.g., AGI and "human-like AI") that are also relative but which index against human performance.

[176] See also Appendix 1D.

[177] An effect reminiscent of the infamous "AI effect," whereby, as John McCarthy famously lamented, "[a]s soon as it works, no one calls it AI anymore." As quoted in: Vardi, Moshe Y. 'Artificial Intelligence: Past and Future'. *Communications of the ACM* 55, no. 1 (January 2012): 5. https://doi.org/10.1145/2063176.2063177.).

[178] I thank Kevin Frazier for observations on this point.

[179] 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. Ftn 6. ("We intentionally speak about 'highly-capable models' instead of 'frontier models'. The 'frontier' refers to the cutting-edge of AI development [18], however the frontier of cutting-edge AI moves forward as AI research progresses. This means that some highly capable systems of concern—those capable of exhibiting dangerous capabilities with the potential to cause significant physical and societal-scale harm—will sit behind the frontier of AI capability. Even if these models are behind the frontier, we should still exercise caution in deciding to release such models, all else being equal.").

III. Defining the advanced AI governance epistemic community

Beyond the object of concern of "advanced AI" (in all its diverse forms), researchers in the emerging field concerned with the impacts and risks of advanced AI systems have begun to specify a range of other terms and concepts, relating to the *tools* for intervening in and on the development of advanced AI systems in socially beneficial ways, terms by which this community's members conceive of the overarching approach or constitution of their *field*, and *theories of change*.

1. Defining the tools for policy intervention

First off, those writing about the risks and regulation of AI have proposed a range of terms in describing the tools, practices, or nature of governance interventions that could be used in response (see Table 7).

Table 7: Definitions of AI strategy, policy, governance

Class	**Terms** [number of definitions surveyed][180]	**Themes**
"Strategy"	→ AI strategy research [1] → AI strategy [1] → Long-term impact strategies [1] → Strategy [1] → AI macrostrategy [1]	→ Focus on big-picture strategic questions to "navigate the transition to a world with advanced AI systems" → Focus on high-level questions to help inform decisions and prioritize use of resources
"Policy"	→ AI policy [1] → AI policymaking strategy [1]	→ Focus on a range of soft and hard governance measures and tools → Organize a research field to inform and shape AI in a responsible, ethical, and robust manner
"Governance"	→ AI governance [5] → Collaborative governance of AI technology [1] → AGI safety and governance practices [1]	→ The study of norms, policies, and institutions to help shape social outcomes from AI systems.

Like the term "advanced AI", these terms set out *objects of study* in scoping the practices or tools of AI governance. They matter insofar as they link these terms to tools for intervention.

[180] For the specific definitions, see Appendix 2A.

Nonetheless, these terms do not capture the methodological dimension of how different approaches to advanced AI governance have approached these issues—nor the normative question of why different research communities have been driven to focus on the challenges from advanced AI in the first place.[181]

2. Defining the field of practice: Paradigms

Thus, we can next consider different ways that practitioners have defined the *field* of advanced AI governance.[182] Researchers have used a range of terms to describe the field of study that focuses on understanding the trajectory to forms of, or impacts of advanced AI and how to shape these. While these have significant overlaps in practice, it is useful to distinguish some key terms or framings of the overall project (Table 8).

Table 8: Definitions of advanced, transformative, and long-term(ist) AI governance

Term [# of definitions][183]	**Focus**
AI governance [3]	→ Governance of AI systems significantly more advanced and capable than those today
Transformative AI governance [1]	→ Governance of societally transformative impacts from AI systems
Longterm(ist) AI governance[184] [4]	→ Governing AI from a perspective that cares about improving the technology's impacts on the future trajectory of society, from (a) various ethical perspectives or (b) a specifically longtermist ethical perspective

[181] This is not to say that building a strong normative commitment into the definition of a research field is without risk. Indeed, in the context of the study of existential risks, Cremer and Kemp have argued that this field should more clearly separate "the science of risk from the moral evaluation of risk," consequently proposing a distinction between: (1) "extinction ethics," which explores the ethical implications of extinction; (2) "existential ethics," which explores the ethical implications of different societal forms and futures in order to understand what (non-extinction) events or trajectories should count as "existential risks" (e.g., dystopias under most value theories), and they argue that the ethical studies of these futures should be separated from (3) the "analysis of human extinction and global catastrophe" per se. Cremer, Carla Zoe, and Luke Kemp. 'Democratising Risk: In Search of a Methodology to Study Existential Risk', 28 December 2021. https://papers.ssrn.com/abstract=3995225. Pg. 12,18. Still, sociologically, it remains important to understand the motives or core concerns of different communities in the advanced AI governance field.

[182] For analogous debates over the appropriate framing of the field focused on technical questions around AI, see previously: Christiano, Paul. 'AI "Safety" vs "Control" vs "Alignment"'. Medium, 19 November 2016. https://ai-alignment.com/ai-safety-vs-control-vs-alignment-2a4b42a863cc. As well as more recently Cotra, Ajeya. '"Aligned" Shouldn't Be a Synonym for "Good"'. Planned Obsolescence, 26 March 2023. https://www.planned-obsolescence.org/aligned-vs-good/.

[183] For the specific definitions, see Appendix 2B.

[184] Generally speaking, these two overlap but are not homologous. Long-term AI governance focuses on governing AI from a perspective that cares about improving the technology's impacts on the future trajectory of society, from various ethical perspectives. Longtermist AI governance is a sub-school or special case, in that it has the same focus on improving AI technology's impacts on the future but more specifically approaches these issues from a longtermist ethical perspective.

However, while these terms show some different focus and emphasis, and different normative commitments, this need not preclude an overall holistic approach. To be sure, work and researchers in this space often hold *diverse expectations* about the trajectory, form, or risks of future AI technologies; diverse *normative* commitments and motivations for studying these; and distinct *research methodologies* given their varied disciplinary backgrounds and epistemic precommitments.[185] However, even so, many of these communities remain united by a shared perception of the technology's stakes—the shared view that shaping the impacts of AI is and should be a significant global priority.[186]

As such, one takeaway here is not that scholars or researchers need pick any one of these approaches or conceptions of the field. Rather, there is a significant need for any advanced AI governance field to maintain a holistic approach, which includes many distinct motivations and methodologies. As suggested by Dafoe,

> "AI governance would do well to emphasize scalable governance: work and solutions to pressing challenges which will also be relevant to future extreme challenges. Given all this potential common interest, the field of AI governance should be inclusive to heterogenous motivations and perspectives. A holistic sensibility is more likely to appreciate that the missing puzzle pieces for any particular challenge could be found scattered throughout many disciplinary domains and policy areas."[187]

In this light, one might consider and frame advanced AI governance as an inclusive and holistic field, concerned with, broadly, "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."

3. Defining theories of change

Finally, researchers in this field have been concerned not just with studying and understanding the strategic parameters of the development of advanced AI systems,[188] but also with considering ways to intervene upon it, given particular assumptions or views about the form, trajectory, societal impacts, or risky capabilities of this technology.

Thus, various researchers have defined terms that aim to capture the connection between immediate interventions or policy proposals, and the eventual goals they are meant to secure (see Table 9).

[185] For broader discussion, see also: Sundaram, Lalitha, Matthijs M. Maas, and S. J. Beard. 'Seven Questions for Existential Risk Studies', (2023 Forthcoming) Managing Extreme Technological Risk (ed. Catherine Rhodes). 25 May 2022. https://doi.org/10.2139/ssrn.4118618.

[186] One version of this is for instance reflected in the 2023 "Statement on AI Risk" by the Center for AI Safety, which reads that "[m]itigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Center for AI Safety. 'Statement on AI Risk', 30 May 2023. https://www.safe.ai/statement-on-ai-risk. See also Dafoe, Allan. 'AI Governance: Opportunity and Theory of Impact', 17 September 2020. https://www.allandafoe.com/opportunity. ("I believe advances in AI are likely to be among the most impactful global developments in the coming decades, and that AI governance will become among the most important global issue areas").

[187] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. 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/9780197579329.013.2. Pg. 3.

[188] For a definition of this term, see also Appendix 3.

Table 9: Definitions of theories of change

Term	Definitions
(Analytic) frame	→ "[A] conceptual orientation that makes salient some aspects of an issue, including cues for what needs to be understood, how to approach the issue, what your goals and responsibilities are, what roles to see yourself as having, what to pay attention to, and what to ignore."[189]
Theory of impact	→ A "simple two stage asset-decision model of research impact"[190] whereby *impactful decisions* will be made by key actors at some future point, and where earlier research can provide *assets* that help these decisions to be made well.[191]
Path to impact[192]	→ "[T]he concrete intervention that future [us] thinks is more valuable than further research—can eventually be tackled with maximum force."[193]
Theory of change	→ "[D]efines long-term goals and then maps backward to identify necessary preconditions [...] explains the process of change by outlining causal linkages in an initiative, i.e., its shorter-term, intermediate, and longer-term outcomes."[194] → "[A] visual depiction of your strategy," linking activities, outcomes, impacts, or goals."[195]
Theory of victory	→ "[C]omplete stories about how humanity successfully navigates the transition to a world with advanced AI."[196]

[189] Stein-Perlman, Zach. 'Framing AI Strategy'. AI Impacts, 6 February 2023. https://aiimpacts.org/framing-ai-strategy/.

[190] Dafoe, Allan. 'AI Governance: Opportunity and Theory of Impact', 17 September 2020. https://www.allandafoe.com/opportunity

[191] Ibid. ("At some point in the causal chain, impactful decisions will be made, be they by AI researchers, activists, public intellectuals, CEOs, generals, diplomats, or heads of state. We want our research activities to provide assets that will help those decisions to be made well. These assets can include: technical solutions; strategic insights; shared perception of risks; a more cooperative worldview; well-motivated and competent advisors; credibility, authority, and connections for those experts. There are different perspectives on which of these assets, and the breadth of the assets, that are worth investing in.").

[192] Garfinkel, Benjamin. 'AI Strategy: Pathways for Impact' (draft shared with author).

[193] Gloor, Lukas. 'Identifying Plausible Paths to Impact and Their Strategic Implications'. *Center on Long-Term Risk* (blog), 14 August 2016. https://longtermrisk.org/identifying-plausible-paths-to-impact/.

[194] Aird, Michael. 'Do Research Organisations Make Theory of Change Diagrams? Should They?' EA Forum, 22 July 2020. https://forum.effectivealtruism.org/posts/LgLYCGCs8Nji3oEWj/do-research-organisations-make-theory-of-change-diagrams. (noting that theories of change can differ in terms of: "Forward chaining vs backward chaining; Self-directed vs other-directed (or proactive vs reactive); Speculative/curiosity-driven vs explicit/foreseeable paths to impact; Fundamental/basic vs applied").

[195] Moss, Ian David. 'A Short Introduction to Theory of Change'. LessWrong, 11 October 2019. https://www.lesswrong.com/posts/xQk6feH9pmy6mKt3x/a-short-introduction-to-theory-of-change.

[196] 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.

Drawing on these terms, one might also articulate new terms that incorporate elements from the above.[197] For instance, one could define a "strategic approach" as a cluster of correlated views on advanced AI governance, encompassing (1) broadly shared *assumptions* about the key technical and governance parameters of the challenge; (2) a broad *theory of victory* and impact story about what solving this problem would look like; (3) a broadly shared *view of history,* with historical analogies to provide comparison, grounding, inspiration, or guidance; and (4) a set of *intermediate strategic goals* to be pursued, giving rise to near-term *interventions* that would contribute to reaching these.

Conclusion

The community focused on governing advanced AI systems has developed a rich and growing body of work. However, it has often lacked clarity, not only regarding many key empirical and strategic questions, but also regarding many of its fundamental terms. This includes different definitions for the relevant object of analysis—that is, species of "advanced AI"—as well as different framings for the instruments of policy, different paradigms or approaches to the field itself, and distinct understandings of what it means to have a theory of change to guide action.

This report has reviewed a range of terms for different analytical categories in the field. It has discussed three different purposes for seeking definitions for core terms, and why and how (under a "regulatory" purpose) the choice of terms matters to both the study and practice of AI governance. It then reviewed analytical definitions of advanced AI used across different clusters which focus on the *forms* or design of advanced AI systems, the (hypothesized) scientific *pathways* towards developing these systems, the technology's broad *societal impacts*, and the specific *critical capabilities* achieved by particular AI systems. The report then briefly reviewed analytical definitions of the tools for intervention, such as "policy" and governance", before discussing definitions of the field and community itself and definitions for theories of change by which to prioritize interventions.

This field of advanced AI governance has shown a penchant for generating many concepts, with many contesting definitions. Of course, while any emerging field will necessarily engage in a struggle to define itself, this field has seen a particularly broad range of terms, perhaps reflecting the disciplinary range. Eventually, the community may need to more intentionally and deliberately commit to some terms. In the meantime, those who engage in debate within and beyond the field should at least have greater clarity about the ways that these concepts are used and understood, and about the (regulatory) implications of some of these terms. This report has aimed to provide such greater clarity in order to help provide greater context for more informed and clear discussions about questions in and around the field.

[197] This has similarities and overlap to the concept of an "(analytical) frame" in: Stein-Perlman, Zach. 'Framing AI Strategy'. AI Impacts, 6 February 2023. https://aiimpacts.org/framing-ai-strategy/. However it is more action-oriented.

Appendix 1: Lists of definitions for advanced AI terms

This appendix provides a detailed list of definitions for advanced AI systems, with sources. These may be helpful for readers to explore work in this field in more detail; to understand the longer history and evolution of many terms; and to consider the strengths and drawbacks of particular terms, and of specific language, for use in public debate, policy formulation, or even in direct legislative texts.

1.A. Definitions focused on the form of advanced AI

Different definitional approaches emphasize distinct aspects or traits that would characterize the form of advanced AI systems—such as that it is "mind-like", performs "autonomously", "is general-purpose", "performs like a human", "performs general-purpose tasks like a human", etc. However, it should be noted that there is significant overlap, and many of these terms are often (whether correctly or not) used interchangeably.[198]

Advanced AI is mind-like & really thinks

→ **Strong AI**

→ An "appropriately programmed computer [that] really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states."[199]

→ "The assertion that machines could possibly act intelligently (or, perhaps better, act as if they were intelligent) is called the 'weak AI' hypothesis by philosophers, and the assertion that machines that do so are actually thinking (as opposed to simulating thinking) is called the 'strong AI' hypothesis."[200]

→ "the combination of Artificial General Intelligence/Human-Level AI and Superintelligence."[201]

Advanced AI is autonomous

→ **Autonomous machine intelligence:** "intelligent machines that learn more like animals and humans, that can reason and plan, and whose behavior is driven by intrinsic objectives, rather than by hard-wired programs, external supervision, or external rewards."[202]

→ **Autonomous artificial intelligence**: "artificial intelligence that can adapt to external environmental challenges. Autonomous artificial intelligence can be similar to animal intelligence, called (specific)

[198] See for example: Hendrycks, Dan, and Mantas Mazeika. 'X-Risk Analysis for AI Research'. arXiv, 21 July 2022. http://arxiv.org/abs/2206.05862. Pg. 36 ("we use the term "strong AI." We use this term synonymously with "AGI" and "human-level AI.").

[199] 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.

[200] Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River: Pearson, 2016. Pg. 1020.

[201] 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.

[202] 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

animal-level autonomous artificial intelligence, or unrelated to animal intelligence, called non-biological autonomous artificial intelligence."[203]

→ **General artificial intelligence**: "broadly capable AI that functions autonomously in novel circumstances".[204]

Advanced AI is human-like

→ **Human-level AI (HLAI)**

→ "systems that operate successfully in the common sense informatic situation [defined as the situation where] the known facts are incomplete, and there is no a priori limitation on what facts are relevant. It may not even be decided in advance what phenomena are to be taken into account. The consequences of actions cannot be fully determined. The common sense informatic situation necessitates the use of approximate concepts that cannot be fully defined and the use of approximate theories involving them. It also requires nonmonotonic reasoning in reaching conclusions."[205]

→ "machines exhibiting true human-level intelligence should be able to do many of the things humans are able to do. Among these activities are the tasks or 'jobs' at which people are employed. I suggest we replace the Turing test by something I will call the 'employment test.' To pass the employment test, AI programs must… [have] at least the potential [to completely automate] economically important jobs."[206]

→ "AI which can reproduce everything a human can do, approximately. [...] [this] can mean either AI which can reproduce a human at any cost and speed, or AI which can replace a human (i.e. is as cheap as a human, and can be used in the same situations.)"[207]

→ "An artificial intelligence capable of matching humans in every (or nearly every) sphere of intellectual activity."[208]

Advanced AI is general-purpose

→ **Foundation model**

[203] 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.

[204] 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.

[205] 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.

[206] 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 (2013) leans towards this as a working operationalization for AGI. Muelhauser, Luke. 'What Is AGI?' Machine Intelligence Research Institute, 11 August 2013. https://intelligence.org/2013/08/11/what-is-agi/.

[207] AI Impacts. 'Human-Level AI'. AI Impacts, 23 January 2014. https://aiimpacts.org/human-level-ai/.

[208] Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 229.

- → "models trained on broad data at scale [...] that are adaptable to a wide range of downstream tasks."[209]
- → "AI systems with broad capabilities that can be adapted to a range of different, more specific purposes. [...] the original model provides a base (hence 'foundation') on which other things can be built."[210]

→ **General-purpose AI systems (GPAIS)**

- → "an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed."[211]
- → "An AI system that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained."[212]
- → "An AI system that can accomplish a range of distinct *valuable* tasks, including some for which it was not specifically trained."[213]
- → See also "general-purpose AI models": "AI models that are designed for generality of their output and have a wide range of possible applications."[214]

→ **Comprehensive AI services (CAIS)**

- → "asymptotically recursive improvement of AI technologies in distributed systems [which] contrasts sharply with the vision of self-improvement internal to opaque, unitary agents. [...] asymptotically comprehensive, superintelligent-level AI services that—crucially—can include the service of developing new services, both narrow and broad, [yielding] a model of flexible,

[209] Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 'On the Opportunities and Risks of Foundation Models'. *arXiv:2108.07258 [Cs]*, 16 August 2021. http://arxiv.org/abs/2108.07258.

[210] Toner, Helen. 'What Are Generative AI, Large Language Models, and Foundation Models?' *Center for Security and Emerging Technology* (blog), 12 May 2023. https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/. See also Jones, Elliot. 'Explainer: What Is a Foundation Model?' Ada Lovelace Institute, 17 July 2023. https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

[211] European Parliament. 'DRAFT Compromise Amendments on the Draft Report Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts', 9 May 2023. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf.; as discussed in: Jones, Elliot. 'Explainer: What Is a Foundation Model?' Ada Lovelace Institute, 17 July 2023. https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

[212] 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.

[213] See Campos, Simeon, and Romain Laurent. 'A Definition of General-Purpose AI Systems: Mitigating Risks from the Most Generally Capable Models'. SSRN Scholarly Paper. Rochester, NY, 19 April 2023. https://papers.ssrn.com/abstract=4423706. (emphasis in original)

[214] Maham, Pegah, and Sabrina Küspert. 'Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks'. Stiftung Neue Verantwortung, July 2023. https://www.stiftung-nv.de/de/publikation/governing-general-purpose-ai-comprehensive-map-unreliability-misuse-and-systemic-risks.

general intelligence in which agents are a class of service-providing products, rather than a natural or necessary engine of progress in themselves."[215]

Advanced AI is general-purpose & of human-level performance

→ **Artificial general intelligence (AGI)** *[task performance definitions]*[216]

→ "systems that exhibit the broad range of general intelligence found in humans."[217]

→ "Artificial intelligence that is not specialized to carry out specific tasks, but can learn to perform as broad a range of tasks as a human."[218]

→ AI systems with "the ability to achieve a variety of goals, and carry out a variety of tasks, in a variety of different contexts and environments."[219]

→ AI systems which "can reason across a wide range of domains, much like the human mind."[220]

→ "machines designed to perform a wide range of intelligent tasks, think abstractly and adapt to new situations."[221]

→ "AI that is capable of solving almost all tasks that humans can solve."[222]

[215] 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.

[216] For a detailed recent review and ontology, see: Morris, Meredith Ringel, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 'Levels of AGI: Operationalizing Progress on the Path to AGI'. arXiv, 4 November 2023. https://doi.org/10.48550/arXiv.2311.02462. For previous reviews, see: Muelhauser, Luke. 'What Is AGI?' Machine Intelligence Research Institute, 11 August 2013. https://intelligence.org/2013/08/11/what-is-agi/. See also 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. 2. [listing definitions focused on "the hypothetical ability of an intelligent agent to understand or learn any intellectual task that a human can," or "the capacity of an engineered system to display the same rough sort of general intelligence as humans," or "the representation of generalized human cognitive abilities in software.", citing sources].

[217] 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.

[218] Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 227.

[219] 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.

[220] 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, pg. 8.

[221] 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.

[222] 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.

→ "AIs that can generalize well enough to produce human-level performance on a wide range of tasks, including abstract low-data tasks."[223]

→ "The AI that [...] can do most everything we humans can do, and possibly much more."[224]

→ "[a]n AI that has a level of intelligence that is either equivalent to or greater than that of human beings or is able to cope with problems that arise in the world that surrounds human beings with a degree of adequacy at least similar to that of human beings.".[225]

→ "an agent that has a world model that's vastly more accurate than that of a human in, at least, domains that matter for competition over resources, and that can generate predictions at a similar rate or faster than a human."[226]

→ "type of AI system that addresses a broad range of tasks with a satisfactory level of performance [or in a stronger sense] systems that not only can perform a wide variety of tasks, but all tasks that a human can perform."[227]

→ "[AI with] cognitive capabilities fully generalizing those of humans."[228]

- See also the subdefinition of autonomous AGI (AAGI) as "an autonomous artificial agent with the ability to do essentially anything a human can do, given the choice to do so—in the form of an autonomously/internally determined directive—and an amount of time less than or equal to that needed by a human."[229]

→ "a machine-based system that can perform the same general-purpose reasoning and problem-solving tasks humans can."[230]

[223] Ngo, Richard. 'AGI Safety From First Principles', 2020. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ. Pg. 5.

[224] Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. Macmillan Publishers, 2019. https://us.macmillan.com/books/9780374715236/artificialintelligence.

[225] Landgrebe, Jobst, and Barry Smith. *Why Machines Will Never Rule the World: Artificial Intelligence without Fear*. 1st edition. Routledge, 2022. pg. xi.

[226] See Ricon, Jose Luis. 'Set Sail For Fail? On AI Risk'. *Nintil*, 4 August 2022. https://nintil.com/ai-safety.

[227] 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.

[228] Critch, Andrew. '"Tech Company Singularities", and Steering Them to Reduce x-Risk'. LessWrong, 13 May 2022. https://forum.effectivealtruism.org/posts/KopQknZEtjZdoGorT/tech-company-singularities-and-steering-them-to-reduce-x.

[229] Ibid.

[230] 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/. For the purposes of question resolution, this definition is operationalized as: "a single unified software system that can satisfy the following criteria, all completable by at least some humans.

- Able to reliably pass a 2-hour, adversarial Turing test during which the participants can send text, images, and audio files (as is done in ordinary text messaging applications) during the course of their conversation. An 'adversarial' Turing test is one in which the human judges are instructed to ask interesting and difficult questions, designed to advantage human participants, and to successfully unmask the computer as an impostor. A single demonstration of an AI passing such a Turing test, or one that is sufficiently similar, will be sufficient for this condition, so long as the test is well-designed to the estimation of Metaculus Admins.

→ "an AI system that equals or exceeds human intelligence in a wide variety of cognitive tasks."[231]

→ "AI systems that achieve or exceed human performance across a wide range of cognitive tasks".[232]

→ "hypothetical type of artificial intelligence that would have the ability to understand or learn any intellectual task that a human being can."[233]

→ "a shorthand for any intelligence [...] that is flexible and general, with resourcefulness and reliability comparable to (or beyond) human intelligence."[234]

→ "systems that demonstrate broad capabilities of intelligence as [...] [a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience], with the additional requirement, perhaps implicit in the work of the consensus group, that these capabilities are at or above human-level."[235]

→ "autonomous artificial intelligence that reaches Human-level intelligence. It can adapt to external environmental challenges and complete all tasks that humans can accomplish, achieving human-level intelligence in all aspects."[236]

- Has general robotic capabilities, of the type able to autonomously, when equipped with appropriate actuators and when given human-readable instructions, satisfactorily assemble a (or the equivalent of a) circa-2021 Ferrari 312 T4 1:8 scale automobile model. A single demonstration of this ability, or a sufficiently similar demonstration, will be considered sufficient.
- High competency at a diverse fields of expertise, as measured by achieving at least 75% accuracy in every task and 90% mean accuracy across all tasks in the Q&A dataset developed by Dan Hendrycks et al.
- Able to get top-1 strict accuracy of at least 90.0% on interview-level problems found in the APPS benchmark introduced by Dan Hendrycks, Steven Basart et al. Top-1 accuracy is distinguished, as in the paper, from top-k accuracy in which k outputs from the model are generated, and the best output is selected.

By 'unified' we mean that the system is integrated enough that it can, for example, explain its reasoning on a Q&A task, or verbally report its progress and identify objects during model assembly."

[231] 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.

[232] 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. Pg. 3.

[233] Definition elicited from ChatGPT (December 5th, 2022).

[234] Marcus, Gary. 'Dear Elon Musk, Here Are Five Things You Might Want to Consider about AGI'. Substack newsletter. *Marcus on AI* (blog), 31 May 2022. https://garymarcus.substack.com/p/dear-elon-musk-here-are-five-things.

[235] 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. Referring to a 1994 definition of intelligence provided in: Gottfredson, Linda S. 'Mainstream Science on Intelligence: An Editorial with 52 Signatories, History, and Bibliography'. *Intelligence* 24, no. 1 (January 1997): 13–23. https://doi.org/10.1016/S0160-2896(97)90011-8.

[236] 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. (noting that "It is also known as Human-Level AI").

→ **Robust artificial intelligence**: "intelligence that, while not necessarily superhuman or self-improving, can be counted on to apply what it knows to a wide range of problems in a *systematic* and *reliable* way, synthesizing knowledge from a variety of sources such that it can reason *flexibly* and *dynamically* about the world, *transferring* what it learns in one context to another, in the way that we would expect of an ordinary adult."[237]

Advanced AI is general-purpose & beyond-human-performance

→ **AI+**: "artificial intelligence of greater than human level (that is, more intelligent than the most intelligent human)"[238]

→ **(Machine/Artificial) superintelligence (ASI)**:

→ "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills."[239]

→ "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest."[240]

→ "an AI significantly more intelligent than humans in all respects."[241]

→ "Artificial intelligence that can outwit humans in every (or almost every) intellectual sphere."[242]

[237] 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.

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

[239] Bostrom, Nick. 'How Long Before Superintelligence?' *International Journal of Futures Studies* 2 (1998). https://nickbostrom.com/superintelligence.

[240] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. Pg. 22. This analysis distinguishes a range of subtypes:

→ **Speed superintelligence**: "A system that can do all that a human intellect can do, but much faster." (ibid. pg 53)
→ **Collective superintelligence**: "A system composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system." (ibid. pg 54)
→ **Quality superintelligence**: "A system that is at least as fast as a human mind and vastly qualitatively smarter." (ibid. pg 56)

This analysis further distinguishes different "castes" of advanced AI:

→ **Oracle**: "a question-answering system."
→ **Genie**: "a command-executing system [which] receives a high-level command, carries it out, then pauses to await the next command."
→ **Sovereign**: "a system that has an open-ended mandate to operate in the world in pursuit of broad and possibly very long-range objectives."
→ **Tool**: "regular software that simply does what it is programmed to do." (ibid. Pg 145–55).

On Tool AI, see also the definition by Karnofsky, Holden. 'Thoughts on the Singularity Institute (SI)'. LessWrong, 2012. https://www.lesswrong.com/posts/6SGqkCgHuNr7d4yJm/thoughts-on-the-singularity-institute-si. ("artificial intelligence that is built to be used as a tool by the creators, rather than being an agent with its own action and goal-seeking behavior").

[241] 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.

[242] Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 231.

→ "future AI systems dramatically more capable than even AGI."[243]

→ "Artificial General Intelligence that has surpassed humans in all aspects of human intelligence."[244]

→ "AI that might be as much smarter than us as we are smarter than insects."[245]

→ See also "machine superintelligence" *[form and impact]*: "general artificial intelligence greatly outstripping the cognitive capacities of humans, and capable of bringing about revolutionary technological and economic advances across a very wide range of sectors on timescales much shorter than those characteristic of contemporary civilization."[246]

→ **Superhuman general-purpose AI (SGPAI)**: "general purpose AI systems [...] that are simultaneously as good or better than humans across nearly all tasks."[247]

→ **Highly capable foundation models**: "Foundation models that exhibit high performance across a broad domain of cognitive tasks, often performing the tasks as well as, or better than, a human."[248]

1.B. Definitions focused on the pathways towards advanced AI

First-principles pathways: "De novo AGI"

Pathways based on new fundamental insights in computer science, mathematics, algorithms, or software, producing advanced AI systems that may, but need not mimic human cognition.[249]

[243] Altman, Sam, Greg Brockman, and Ilya Sutskever. 'Governance of Superintelligence'. OpenAI, 22 May 2023. https://openai.com/blog/governance-of-superintelligence.

[244] 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.

[245] Chapman, David. *Better without AI*, 2023. https://betterwithout.ai/. (chapter: "what is the Scary kind of AI?").

[246] 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. Note, this definition of "machine superintelligence" mixes elements of different definitional approaches, by specifying not just its anticipated *form* but also its anticipated *societal impact*.

[247] 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. (clarifying that "This naming is used to emphasis that generality and capability are distinct. General-purpose AI is here, and likely to simply get more powerful; different adjectives like 'human-competitive' and 'superhuman' in this essay will indicate levels of capability we can expect to move through. We should not necessarily expect some new breakthrough or step-change to something fundamentally different and worth calling 'AGI.'").

[248] 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.

[249] 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.

→ **De novo AGI**: "AGI built from the ground up."[250]

Scaling pathways: "Prosaic AGI", "frontier (AI) model" [compute threshold]

Approaches based on "brute forcing" advanced AI,[251] by running (one or more) existing AI approaches (such as transformer-based LLMs)[252] with increasingly more computing power and/or training data, as per the "scaling hypothesis."[253]

→ **Prosaic AGI**: AGI "which can replicate human behavior but doesn't involve qualitatively new ideas about 'how intelligence works.'"[254]

→ **Frontier (AI) model** *[compute threshold]*:[255]

→ "foundation model that is trained with more than some amount of computational power—for example, 10^26 FLOP."[256]

→ "models within one order of magnitude of GPT-4 (>2e24 FLOP)."[257]

Evolutionary pathways: "[AGI] from evolution"

Approaches based on algorithms competing to mimic the evolutionary brute search process that produced human intelligence.[258]

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").

[250] 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.

[251] 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.

[252] Strictly speaking, the scaling paradigm could be one applied to or combined with most of the other "pathways." In practice, it has often been increasingly applied especially to approaches that expect that training existing neural network algorithms on more data and with more compute is a path towards general intelligence. See Branwen, Gwern. 'The Scaling Hypothesis', 28 May 2020. https://www.gwern.net/Scaling-hypothesis. On the appearance of empirical scaling laws in neural network-based large language models, see: 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.

[253] Branwen, Gwern. 'The Scaling Hypothesis', 28 May 2020. https://www.gwern.net/Scaling-Hypothesis.

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

[255] Note: the compute-threshold approach to defining "frontier AI" is primarily a *development*-path-based definition (focusing as it does on how the system is defined in relation to the *scaling* approach), it could also be considered a *critical-capability*-focused definition, because the motivation behind operationalizing the concept in this way has often been for a particular compute threshold to serve as a proxy for particular (unpredictable) dangerous capabilities.

[256] 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'. Pg. 35.

[257] Metaculus. 'Will There Be a Frontier AI Lab Outside the US before 2026?' Accessed 4 October 2023. https://manifold.markets/MetaculusBot/will-there-be-a-frontier-ai-lab-out.

[258] 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

→ **[AGI] from evolution**: "[AGI re-evolved through] genetic algorithms on computers that are sufficiently fast to recreate on a human timescale the same amount of cumulative optimization power that the relevant processes of natural selection instantiated throughout our evolutionary past."[259]

Reward-based pathways: "[AGI] from powerful reinforcement learning agents", "powerful deep learning models"

Approaches based on running reinforcement learning systems with simple rewards in rich environments.

→ **[AGI] from powerful reinforcement learning agents**: "powerful reinforcement learning agents, when placed in complex environments, will in practice give rise to sophisticated expressions of intelligence."[260]

→ **Powerful deep learning models**: "a powerful neural network model [trained] to simultaneously master a wide variety of challenging tasks (e.g. software development, novel-writing, game play, forecasting, etc.) by using reinforcement learning on human feedback and other metrics of performance."[261]

Bootstrapping pathways:[262] "Seed AI"

Approaches that pursue a minimally intelligent core system capable of subsequent recursive (self)-improvement,[263] potentially leveraging hardware or data "overhangs."[264]

→ **Seed AI**:

→ "an AI designed for self-understanding, self-modification, and recursive self-improvement."[265]

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.

[259] 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

[260] 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.

[261] 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. For more discussion, see also Cotra, Ajeya. 'Supplement to "Why AI Alignment Could Be Hard"'. Cold Takes, 19 September 2021. https://www.cold-takes.com/supplement-to-why-ai-alignment-could-be-hard/. Cotra and Karnofsky have called this training approach Human Feedback on Diverse Tasks (HFDT).

[262] For an early overview, see: 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. Note, there is some overlap between this category and the "exponential" capabilities category.

[263] For a different but related framework, see also "Experience-based AI" (EXPAI), which focuses on shaping a system's growth towards a robust and trustworthy, benevolent, and well-behaved agent. Steunebrink, Bas R., Kristinn R. Thórisson, and Jürgen Schmidhuber. 'Growing Recursive Self-Improvers'. In *Artificial General Intelligence*, edited by Bas Steunebrink, Pei Wang, and Ben Goertzel, 9782:129–39. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-41649-6_13.

[264] Notably, cases where self-improvement comes about as a result of the utilization of existing hardware or data overhangs may be more surprising or sudden than those that involve self-improvement. I thank John-Clark Levin on this point.

[265] 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.

→ "a strongly self-improving process, characterized by improvements to the content base that exert direct positive feedback on the intelligence of the underlying improving process."[266]

→ "The first AI in a series of recursively self-improving systems."[267]

Neuro-inspired pathways: "NeuroAI", "brain-like-AGI", "neuromorphic AI"

Various forms of biologically-inspired, brain-inspired,[268] or brain-imitative approaches that draw on neuroscience and/or "connectomics" to reproduce general intelligence.

→ **NeuroAI**: "[a field] at the intersection of neuroscience and AI, [that] is based on the premise that a better understanding of neural computation will reveal fundamental ingredients of intelligence [...] [which will] lead to artificial agents with capabilities that match those of humans."[269]

→ **Brain-like-AGI**: "algorithms with big-picture similarity to key ingredients of human intelligence, presumably (though not necessarily) as a result of future people reverse-engineering those aspects of the human brain."[270]

→ **Neuromorphic AI**: "AGI based loosely on the principles of the human brain."[271]

Neuro-emulated pathways: "Whole-brain-emulation" (WBE)

Approaches that aim to digitally simulate or recreate the states of human brains at fine-grained level.

→ **Whole-brain-emulation** (WBE):

[266] Ibid. Yudkowsky (2007), pg. 102.

[267] Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 230.

[268] 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 (arguing that "an AI system is biologically inspired when its architecture and functioning include biological constraints that make specific parts of the system biologically realistic. Importantly, a biologically inspired AI system does not necessarily fully emulate or replicate the reference biological system, since different levels of biological realism are possible"). I thank Carla Zoe Cremer for this suggestion.

[269] 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

[270] Byrnes, Steven. '[Intro to Brain-like-AGI Safety] 1. What's the Problem & Why Work on It Now?' AI Alignment Forum, 26 January 2022. https://www.alignmentforum.org/posts/4basF9w9jaPZpoC8R/intro-to-brain-like-agi-safety-1-what-s-the-problem-and-why.

[271] 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. However, note that this usage may be nonstandard, as the term "neuromorphic" is more commonly used to brain-inspired computing *hardware*, not software; see Reuther, Albert, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, and Jeremy Kepner. 'AI and ML Accelerator Survey and Trends'. In *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, 1–10, 2022. https://doi.org/10.1109/HPEC55821.2022.9926331. Pg. 6. ("neuromorphic computing is the research, design, and development of computational hardware that models functionality and processes in brains, including chemical processes and electrical processes").

→ "software (and possibly dedicated non-brain hardware) that models the states and functional dynamics of a brain at a relatively fine-grained level of detail."[272]

→ "The process of making an exact computer-simulated copy of the brain of a particular animal (e.g., a particular human)."[273]

→ **Digital people** *[emulation definition]*: "a computer simulation of a specific person, in a virtual environment [...] perhaps created via mind uploading (simulating human brains) [or] entities unlike us in many ways, but still properly thought of as 'descendants' of humanity."[274]

→ See also related terms: "Ems"[275] or "uploads".

Neuro-integrationist pathways: "Brain-computer-interfaces" (BCI)

Approaches to create advanced AI, based on merging components of human and digital cognition.

→ **Brain-computer-interfaces (BCI)**: "use brain-computer interfaces to position both elements, human and machine, to achieve (or overachieve) human goals."[276]

Embodiment pathways:[277] "Embodied agent"

Based on providing the AI system with a robotic physical "body" to ground cognition and enable it to learn from direct experience of the world.[278]

→ "an embodied agent (e.g., a robot) which learns, through interaction and exploration, to creatively solve challenging tasks within its environment."[279]

[272] 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.

[273] Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 232.

[274] Karnofsky, Holden. 'Digital People Would Be An Even Bigger Deal'. Cold Takes, 27 July 2021. https://www.cold-takes.com/how-digital-people-could-change-the-world/. Elsewhere, however, Karnofsky defines this term with greater reference to its critical (moral) capabilities and clarifies that while emulation would be a most obvious route, there could be distinct pathways towards such entities. Karnofsky, Holden. 'Digital People FAQ'. Cold Takes, 27 July 2021. https://www.cold-takes.com/digital-people-faq/. ("A mind upload would be one form of digital person [...] Mind uploads are the most easy-to-imagine version of digital people, and I focus on them when I talk about why I think digital people will someday be possible and why they would be conscious like we are. But I could also imagine a future of 'digital people' that are not derived from copying human brains, or even all that similar to today's humans.").

[275] Hanson, Robin. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford University Press, 2016. https://doi.org/10.1093/oso/9780198754626.001.0001.; see also Hanson, Robin. 'Whole Brain Emulation - Envisioning Economies And Societies of Emulated Minds'. *PSW Science* (blog), 2 December 2012. https://pswscience.org/meeting/whole-brain-emulation-envisioning-economies-and-societies-of-emulated-minds/.

[276] Ibid. pg. 5.

[277] See for instance 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

[278] 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.

[279] 'Embodied AI Workshop', 2022. https://embodied-ai.org/#organizers.

Modular cognitive architecture pathways

Used in various fields, including in robotics, where researchers integrate well-tested but distinct state-of-the-art modules (perception, reasoning, etc.) to improve agent performance without independent learning.[280]

→ No clear single term.

Hybrid pathways

Approaches that rely on combining deep neural network-based approaches to AI with other paradigms (such as symbolic AI).

→ **Hybrid AI**: "hybrid, knowledge-driven, reasoning-based approach, centered around cognitive models."[281]

1.C. Definitions focused on the aggregate societal impacts of advanced AI

(Strategic) general-purpose technology (GPT)

→ "[AI systems] having an unusually broad and deep impact on the world, comparable to that of electricity, the internal combustion engine, and computers."[282]

→ This has been further operationalized as: "[this] need not emphasize only agent-like AI or powerful AI systems, but instead can examine the many ways even mundane AI could transform fundamental parameters in our social, military, economic, and political systems, from developments in sensor technology, digitally mediated behavior, and robotics. AI and associated technologies could dramatically reduce the labor share of value and increase

[280] See for instance the approach based on the "Common Model of Cognition": West, Robert L. 'Introduction to Volume 1(2)'. *Common Model of Cognition Bulletin* 1, no. 2 (24 June 2020). https://ojs.library.carleton.ca/index.php/cmcb/article/view/2703. I thank José Hernández-Orallo for this suggestion.

[281] 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.

[282] 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. Pg.38. Compare this against the original definitions of GPTs, in: Bresnahan, Timothy, and Manuel Trajtenberg. 'General Purpose Technologies "Engines of Growth"?' *Journal of Econometrics* 65, no. 1 (1995): 83–108. https://econpapers.repec.org/article/eeeeconom/v_3a65_3ay_3a1995_3ai_3a1_3ap_3a83-108.htm ("[technologies] characterised by pervasiveness, inherent potential for technical improvements, and 'innovational complementarities', giving rise to increasing returns-to-scale."); Lipsey, Richard, Kenneth I. Carlaw, and Clifford T. Bekar. 'Economic Transformations: General Purpose Technologies and Long-Term Economic Growth'. OUP Catalogue. Oxford University Press, 2005. https://econpapers.repec.org/bookchap/oxpobooks/9780199290895.htm. ("a technology that initially has much scope for improvement and eventually comes to be widely used, to have many uses, and to have many spillover effects"). For a discussion of AI as a GPT in an economic context, see: Trajtenberg, Manuel. 'AI as the next GPT: A Political-Economy Perspective'. Working Paper. National Bureau of Economic Research, January 2018. https://doi.org/10.3386/w24245. For a historical analysis, see: Crafts, Nicholas. 'Artificial Intelligence as a General-Purpose Technology: An Historical Perspective'. *Oxford Review of Economic Policy* 37, no. 3 (1 September 2021): 521–36. https://doi.org/10.1093/oxrep/grab012. And also: 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, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.5. 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.

inequality, reduce the costs of surveillance and repression by authorities, make global market structure more oligopolistic, alter the logic of the production of wealth, shift military power, and undermine nuclear stability."[283]

→ See also strategic general-purpose technology: "A general purpose technology which has the potential to deliver vast economic value and substantially affect national security, and is consequently of central political interest to states, firms, and researchers."[284]

General-purpose military transformation (GMT)

→ The process by which general-purpose technologies (such as electricity and AI) "influence military effectiveness through a protracted, gradual process that involves a broad range of military innovations and overall industrial productivity growth."[285]

Transformative AI (TAI):[286]

→ "potential future AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution."[287]

→ "AI powerful enough to bring us into a new, qualitatively different future."[288]

→ "software which causes a tenfold acceleration in the rate of growth of the world economy (assuming that it is used everywhere that it would be economically profitable to use it)."[289]

[283] Dafoe, Allan. 'AI Governance: Opportunity and Theory of Impact', 17 September 2020. https://www.allandafoe.com/opportunity.

[284] 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. 35

[285] Ding, Jeffrey, and Allan Dafoe. 'Engines of Power: Electricity, AI, and General-Purpose Military Transformations'. *ArXiv:2106.04338 [Econ, q-Fin]*, 8 June 2021. http://arxiv.org/abs/2106.04338. Pg. 2.

[286] For a review of definitions, see also: Gruetzemacher, Ross, and Jess Whittlestone. 'The Transformative Potential of Artificial Intelligence'. *Futures* 135 (2022): 102884. https://doi.org/10.1016/j.futures.2021.102884.

[287] Karnofsky, Holden. 'Some Background on Our Views Regarding Advanced Artificial Intelligence'. Open Philanthropy Project, 6 May 2016. https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence; see also Muelhauser, Luke. 'What Open Philanthropy Means by "Transformative AI"'. Google Docs, June 2019. https://docs.google.com/document/d/15siOkHQAoSBl_Pu85UgEDWfmvXFotzub31ow3A11Xvo/edit?usp=embed_facebook.; for related formulations that index TAI's impact with reference to the industrial revolution, see also: Cotra, Ajeya. 'Forecasting TAI with Biological Anchors (Draft)'. Open Philanthropy Project, July 2020. https://drive.google.com/drive/folders/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP. Part 1. Pg. 18. ("'software' (i.e. a computer program or collection of computer programs) that has at least as profound an impact on the world's trajectory as the Industrial Revolution did;"); 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. ("advanced AI systems whose long-term impacts may be as profound as the industrial revolution"); Muelhauser, Luke. 'Our AI Governance Grantmaking so Far'. Open Philanthropy, 16 December 2020. https://www.openphilanthropy.org/blog/ai-governance-grantmaking. ("Software that has at least as profound an impact on the world's trajectory as the Industrial Revolution did").

[288] 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/.

[289] Cotra, Ajeya. 'Forecasting TAI with Biological Anchors (Draft)'. Open Philanthropy Project, July 2020. https://drive.google.com/drive/folders/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP. Part 1. Pg. 18.

→ "AI that can go beyond a narrow task ... but falls short of achieving superintelligence."[290]

→ "a range of possible advances with potential to impact society in significant and hard-to-reverse ways."[291]

→ "Any AI technology or application with potential to lead to practically irreversible change that is broad enough to impact most important aspects of life and society."[292]

Radically transformative AI (RTAI)

→ "any AI technology or application which meets the criteria for TAI, and with potential impacts that are extreme enough to result in radical changes to the metrics used to measure human progress and well-being, or to result in reversal of societal trends previously thought of as practically irreversible. This indicates a level of societal transformation equivalent to that of the agricultural or industrial revolutions."[293]

AGI [economic competitiveness definition]

→ "highly autonomous systems that outperform humans at most economically valuable work."[294]

→ "AI systems that power a comparably profound transformation (in economic terms or otherwise) as would be achieved in [a world where cheap AI systems are fully substitutable for human labor]."[295]

[290] Horowitz, Michael C. 'Artificial Intelligence, International Competition, and the Balance of Power'. *Texas National Security Review*, 15 May 2018. https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/.

[291] 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 pg. 100. (giving as examples that "future machine learning systems could be used to optimise management of safety-critical infrastructure […]. Advanced language models could be used in ways that corrupt our online information ecosystem [...] and future advances in AI systems could trigger widespread labour automation").

[292] Gruetzemacher, Ross, and Jess Whittlestone. 'The Transformative Potential of Artificial Intelligence'. *Futures* 135 (2022): 102884. https://doi.org/10.1016/j.futures.2021.102884 pg. 9.

[293] ibid.

[294] OpenAI. 'OpenAI Charter', 9 April 2018. https://openai.com/charter. See also the more general formulation in Altman, Sam. 'Planning for AGI and Beyond'. OpenAI, 24 February 2023. https://openai.com/blog/planning-for-agi-and-beyond ("AI systems that are generally smarter than humans").

[295] Beckstead, Nick, Leopold Aschenbrenner, William MacAskill, Ketan Rama, and Avital Balwit. 'Announcing the Future Fund's AI Worldview Prize'. Effective Altruism Forum, 23 September 2022. https://forum.effectivealtruism.org/posts/W7C5hwq7sjdpTdrQF/announcing-the-future-fund-s-ai-worldview-prize. The full definition and operationalization given is:

"Imagine a world where cheap AI systems are fully substitutable for human labor. E.g., for any human who can do any job, there is a computer program (not necessarily the same one every time) that can do the same job for $25/hr or less. This includes entirely AI-run companies, with AI managers and AI workers and everything being done by AIs.

- How large of an economic transformation would follow? Our guess is that it would be pretty large (see Aghion et al 2017, this post, and Davidson 2021), but—to the extent it is relevant—we want people competing for this prize to make whatever assumptions seem right to them.

For purposes of our definitions, we'll count it as AGI being developed if there are AI systems that power a comparably profound transformation (in economic terms or otherwise) as would be achieved in such a world. Some caveats/clarifications worth noticing:

- A comparably large economic transformation could be achieved even if the AI systems couldn't substitute for literally 100% of jobs, including providing emotional support. E.g., Karnofsky's notion of PASTA would probably count (though that is an empirical question), and possibly some other things would count as well.

→ "future machines that could match and then exceed the full range of human cognitive ability across all economically valuable tasks."[296]

Machine superintelligence [form & impact definition]

→ "general artificial intelligence greatly outstripping the cognitive capacities of humans, and capable of bringing about revolutionary technological and economic advances across a very wide range of sectors on timescales much shorter than those characteristic of contemporary civilization"[297]

1.D. Definitions focused on critical capabilities of advanced AI systems

Systems with critical moral and/or philosophical capabilities

→ **Artificial/Machine consciousness**:

→ "machines that genuinely exhibit conscious awareness."[298]

→ "Weakly construed, the possession by an artificial intelligence of a set of cognitive attributes that are associated with consciousness in humans, such as awareness, self-awareness, or cognitive integration. Strongly construed, the possession by an AI of properly phenomenological states, perhaps entailing the capacity for suffering."[299]

→ **Digital minds**: "machine minds with conscious experiences, desires, and capacity for reasoning and autonomous decision-making [...] [which could] enjoy moral status, i.e. rather than being mere tools of humans they and their interests could matter in their own right."[300]

- If weird enough things happened, the metric of GWP might stop being indicative in the way it normally is, so we want to make sure people are thinking about the overall level of weirdness rather than being attached to a specific measure or observation. E.g., causing human extinction or drastically limiting humanity's future potential may not show up as rapid GDP growth, but automatically counts for the purposes of this definition."

[296] Beniach, Nathan. 'State of AI Report 2023'. Air Street Capital, 12 October 2023. https://www.stateof.ai/. Pg. 5.

[297] 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.

[298] Reggia, James A. 'The Rise of Machine Consciousness: Studying Consciousness with Computational Models'. *Neural Networks* 44 (1 August 2013): 112–31. https://doi.org/10.1016/j.neunet.2013.03.011. For a more recent discussion of what different competing theories of consciousness may tell us about the prospects or feasibility of conscious AI systems, see: Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, et al. 'Consciousness in Artificial Intelligence: Insights from the Science of Consciousness'. arXiv, 22 August 2023. https://doi.org/10.48550/arXiv.2308.08708.

[299] Shanahan, Murray. *The Technological Singularity*. MIT Press Essential Knowledge Series. MIT Press, 2015. https://mitpress.mit.edu/9780262527804/the-technological-singularity/. Pg. 227.

[300] Shulman, Carl, and Nick Bostrom. 'Sharing the World with Digital Minds'. In *Rethinking Moral Status*, edited by Steve Clarke, Hazem Zohny, and Julian Savulescu. Oxford University Press, 2021. https://academic.oup.com/book/41245/chapter/350760172.; see also: Bostrom, Nick, and Carl Shulman. 'Propositions Concerning Digital Minds and Society', 2022, 20. https://www.nickbostrom.com/propositions.pdf ; for an older discussion, see Moravec's "mind children"; Moravec, H. *Mind Children: The Future of Robot and Human Intelligence*. New Ed edition. Cambridge: Harvard University Press, 1990.

→ **Digital people** *[capability definition]*: "any digital entities that (a) had moral value and human rights, like non-digital people; (b) could interact with their environments with equal (or greater) skill and ingenuity to today's people."[301]

→ **Sentient artificial intelligence:** "artificial intelligence (capable of feeling pleasure and pain)."[302]

→ **Robot rights catastrophe:** The point where AI systems are sufficiently advanced that "some people reasonably regard [them] as deserving human or humanlike rights. [while] Other people will reasonably regard these systems as wholly undeserving of human or humanlike rights. [...] Given the uncertainties of both moral theory and theories about AI consciousness, it is virtually impossible that our policies and free choices will accurately track the real moral status of the AI systems we create. We will either seriously overattribute or seriously underattribute rights to AI systems—quite possibly both, in different ways. Either error will have grave moral consequences, likely at a large scale. The magnitude of the catastrophe could potentially rival that of a world war or major genocide."[303]

→ **(Negative) synthetic phenomenology**: "machine consciousness [that] will have preferences of their own, that [...] will autonomously create a hierarchy of goals, and that this goal hierarchy will also become a part of their phenomenal self-model [...] [such that they] will be able to consciously suffer,"[304] creating a risk of an "explosion of negative phenomenology" (ENP) ("Suffering explosion").[305]

→ **Suffering risks**: "[AI that brings] about severe suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far."[306]

→ **Adversarial technological maturity**:

[301] Karnofsky, Holden. 'Digital People FAQ'. Cold Takes, 27 July 2021. https://www.cold-takes.com/digital-people-faq/. For commentary and questions, see: Long, Robert. 'Digital People: Biology versus Silicon'. Substack newsletter. *Experience Machines*, 2 August 2022. https://experiencemachines.substack.com/p/digital-people-biology-versus-silicon.

[302] 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.

[303] Schwitzgebel, Eric. 'The Coming Robot Rights Catastrophe'. *Blog of the APA* (blog), 12 January 2023. https://blog.apaonline.org/2023/01/12/the-coming-robot-rights-catastrophe/. Note, this case highlights that it is not strictly speaking necessary for advanced AI systems to genuinely achieve (morally relevant) traits such as sentience or consciousness in order for them to create impacts that are morally or philosophically disruptive (in a sociological sense). After all, it is exactly in situations where the precise nature of AI system's cognition is unclear, such that precise attribution or determination of moral status remains difficult, that there can be significant risks of societal upheaval over the question of whether to extend legal protections, with risks of inadvertent but catastrophic over- or under-attribution of such status.

[304] 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. 1–2.

[305] Ibid. pg. 3

[306] 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. Pg. 389. See also Tomasik, Brian. 'Astronomical Suffering from Slightly Misaligned Artificial Intelligence', 2018. https://reducing-suffering.org/near-miss/. As well as: 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. Pg. 46.

  - → "the point where there are digital people and/or (non-misaligned) AIs that can copy themselves without limit, and expand throughout space [creating] intense pressure to move - and multiply (via copying) - as fast as possible in order to gain more influence over the world."[307]

  - → "a world in which highly advanced technology has already been developed, likely with the help of AI, and different coalitions are vying for influence over the world."[308]

Systems with critical economic capabilities[309]

- → **High-level machine intelligence (HLMI)**:

  - → "unaided machines [that] can accomplish every task better and more cheaply than human workers."[310]

  - → "an AI system (or collection of AI systems) that performs at the level of an average human adult on key cognitive measures required for economically relevant tasks."[311]

  - → "The spectrum of advanced AI capabilities from next-generation AI systems to artificial general intelligence (AGI). Often used interchangeably with advanced AI."[312]

- → **Tech company singularity**: "a transition of a technology company into a fully general tech company [defined as] a technology company with the ability to become a world-leader in essentially any industry sector, given the choice to do so—in the form of agreement among its Board and CEO—with around one year of effort following the choice."[313]

- → **Artificial capable intelligence (ACI)**:

  - → "AI [that] can achieve complex goals and tasks with minimal oversight."[314]

[307] Karnofsky, Holden. 'How to Make the Best of the Most Important Century?' Cold Takes, 14 September 2021. https://www.cold-takes.com/making-the-best-of-the-most-important-century/.

[308] Ibid.

[309] Discussions on economic impacts often turn on whether the performance of AI systems is anticipated to match (and therefore potentially replace) (1) average human performance or (2) top human performance, or exceed (3) any human performance. I thank John-Clark Levin for pointing out this distinction.

[310] 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 (31 July 2018): 729–54. https://doi.org/10.1613/jair.1.11222. See also the updated survey: 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/. Note: this definition has strong overlaps with the definition of "human-level machine intelligence/AI," though it is more focused on the economic benchmark.

[311] 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 pg. 105.

[312] Kilian, Kyle A. 'From Deus Ex Machina to Society of Mind: The Differential Risk from High-Level Machine Intelligence and the Question of Control'. National Intelligence University, 2022. https://drive.google.com/file/d/17-UPR7rUreW9LHb2ZsqD3hgIXLMi1RuY/view?usp=sharing&usp=embed_facebook. Pg. x.

[313] Critch. '"Tech Company Singularities", and Steering Them to Reduce x-Risk'. EA Forum, 2022. https://forum.effectivealtruism.org/posts/KopQknZEtjZdoGorT/tech-company-singularities-and-steering-them-to-reduce-x.

[314] Suleyman, Mustafa, and Michael Bhaskar. *The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma*. New York: Crown, 2023. Pg. 82. In this proposal, ACI is measured through a "Modern Turing Test,"

- → "a fast-approaching point between AI and AGI: ACI can achieve a wide range of complex tasks but is still a long way from being fully general."[315]

Systems with critical legal capabilities

- → **Advanced artificial judicial intelligence (AAJI)**: "an artificially intelligent system that matches or surpasses human decision-making in all domains relevant to judicial decision-making."[316]
- → **Technological-legal lock-in**: "hybrid human/AI judicial systems [which] risk fostering legal stagnation and an attendant loss of judicial legitimacy."[317]
- → **Legal singularity**: "when the accumulation of a massive amount of data and dramatically improved methods of inference make legal uncertainty obsolete. The legal singularity contemplates complete law. [...] the elimination of legal uncertainty and the emergence of a seamless legal order, which is universally accessible in real time."[318]

Systems with critical scientific capabilities

- → **Process-automating science and technology (PASTA)**: "AI systems that can essentially automate all of the human activities needed to speed up scientific and technological advancement."[319]
- → **Scientist model**: "a single unified transformative model [...] which has flexible general-purpose research skills."[320]

within which an AI would be able to successfully act on the instruction "Go make $1 million on a retail web platform in a few months with just a $100,000 investment." ibid. Pg. 78. See also Suleyman, Mustafa. 'My New Turing Test Would See If AI Can Make $1 Million'. MIT Technology Review, 14 July 2023. https://www.technologyreview.com/2023/07/14/1076296/mustafa-suleyman-my-new-turing-test-would-see-if-ai-can-make-1-million/.

[315] Ibid. pg. 11.

[316] Winter, Christoph, Nick Hollman, and David Manheim. 'Value Alignment for Advanced Artificial Judicial Intelligence'. American Philosophical Quarterly, 19 October 2022. https://papers.ssrn.com/abstract=4252645, quoting 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.

[317] Crootof, Rebecca. '"Cyborg Justice" and the Risk of Technological-Legal Lock-In'. *Columbia Law Review Forum* 119 (5 October 2019): 1–19. https://columbialawreview.org/content/cyborg-justice-and-the-risk-of-technological-legal-lock-in/ Pg. 4.

[318] Alarie, Benjamin. 'The Path of the Law: Towards Legal Singularity'. *University of Toronto Law Journal* 66, no. 4 (1 January 2016): 443–55. https://doi.org/10.3138/UTLJ.4008. For a discussion of what such capabilities could mean for established theories of law, see Sheppard, Brian. 'Warming up to Inscrutability: How Technology Could Challenge Our Concept of Law'. *University of Toronto Law Journal* 68, no. supplement 1 (January 2018): 36–62. https://doi.org/10.3138/utlj.2017-0053.

[319] Karnofsky, Holden. 'Forecasting Transformative AI, Part 1: What Kind of AI?' Cold Takes, 10 August 2021. https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/.

[320] 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.

Systems with critical strategic or military capabilities[321]

→ **Decisive strategic advantage**: "a position of strategic superiority sufficient to allow an agent to achieve complete world domination."[322]

→ **Singleton**: [AI capabilities sufficient to support] "a world order in which there is a single decision-making agency at the highest level."[323]

Systems with critical political capabilities

→ **Stable totalitarianism**: "AI [that] could enable a relatively small group of people to obtain unprecedented levels of power, and to use this to control and subjugate the rest of the world for a long period of time (e.g. via advanced surveillance)."[324]

→ **Value lock-in**:

→ "an event [such as the use of AGI] that causes a single value system, or set of value systems, to persist for an extremely long time."[325]

→ "AGI [that] would make it technologically feasible to (i) perfectly preserve nuanced specifications of a wide variety of values or goals far into the future, and (ii) develop AGI-based institutions that would (with high probability) competently pursue any such values for at least millions, and plausibly trillions, of years."[326]

→ **Actually existing AI (AEAI)**: A paradigm by which the broader ecosystem of AI development, on current trajectories, may produce harmful political outcomes, because "AI as currently funded, constructed, and concentrated in the economy—is misdirecting technological resources towards

[321] Note, this does not review more specific applications of AI systems in military roles, as reflected in concepts concept such as "advanced military AI", an "AI commander", or a "fog-of-war machine" optimized for military deception. Though for literature on these topics, see respectively (on advanced military AI): Maas, Matthijs, Kayla Lucero-Matteucci, and Di Cooke. '10. Military Artificial Intelligence as a Contributor to Global Catastrophic Risk'. In *The Era of Global Risk*, 237–84. Open Book Publishers, 2023. https://www.openbookpublishers.com/books/10.11647/obp.0336/chapters/10.11647/obp.0336.10.; Turchin, Alexey, and David Denkenberger. 'Military AI as a Convergent Goal of Self-Improving AI'. In *Artificial Intelligence Safety and Security*, edited by Roman Yampolskiy. Louiswille: CRC Press, 2018. https://philpapers.org/rec/TURMAA-6. (on an AI commander:) Johnson, James. *The AI Commander: Centaur Teaming, Command, and Ethical Dilemmas*. Oxford, New York: Oxford University Press, 2024 forthcoming; (on Fog-Of-War Machines:) Geist, Edward. 'Fog-of-War Machines'. In *Deterrence under Uncertainty:Artificial Intelligence and Nuclear Warfare*, edited by Edward Geist, 0. Oxford University Press, 2023. https://doi.org/10.1093/oso/9780192886323.003.0006.

[322] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. Pg. 78.

[323] Bostrom, Nick. 'What Is a Singleton?' *Linguistic and Philosophical Investigations* 5, no. 2 (2006): 48–54. https://www.fhi.ox.ac.uk/wp-content/uploads/singleton.pdf

[324] 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.

[325] MacAskill, William. *What We Owe the Future*. New York, NY: Basic Books, 2022. Pg. 78. Compare this also with the definition of "value lock-in" in: 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. Pg. 4 ("[a point in time when] we will invent a technology that will enable the agents alive at that time to maintain their values indefinitely into the future, controlling the broad sweep of the entire rest of the future of civilisation").

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

unproductive and dangerous outcomes. It is driven by a wasteful imitation of human comparative advantages and a confused vision of autonomous intelligence, leading it toward inefficient and harmful centralized architectures."[327]

Systems with critical exponential capabilities

→ **Intelligence explosion**:[328]

→ "explosion to ever greater levels of intelligence, as each generation of machines creates more intelligent machines in turn."[329]

→ "a chain of events by which human-level AI leads, fairly rapidly, to intelligent systems whose capabilities far surpass those of biological humanity as a whole."[330]

→ **Autonomous replication in the real world**: "A model that is unambiguously capable of replicating, accumulating resources, and avoiding being shut down in the real world indefinitely, but can still be stopped or controlled with focused human intervention."[331]

→ **Autonomous AI research**: "A model for which the weights would be a massive boost to a malicious AI development program (e.g. greatly increasing the probability that they can produce systems that meet other criteria for [AI Safety Level]-4 in a given timeframe)."[332]

→ **Duplicator**: [digital people or particular forms of advanced AI that would allow] "the ability to make instant copies of people (or of entities with similar capabilities) [leading to] explosive productivity."[333]

[327] 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. Pg. 1. In their program, this trajectory or program is contrasted to "actually existing digital plurality" (AEDP). Compare this also with the suggested turn away from "machine intelligence," and towards a program of (configuring AI technologies to serve) "machine usefulness" (MU), in Acemoglu, Daron, and Simon Johnson. *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. New York: Public Affairs, 2023. (pg. 316–332).

[328] For the original term, see: Good, I.J. 'Speculations Concerning the First Ultraintelligent Machine'. In *Advances in Computers*, edited by Franz L. Alt and Moris Rubinoff, 6:31–88. New York: Academic Press, 1964. Though note that an earlier reference to a machine-induced "explosion" can already be found in a 1959 lecture: Good, I.J. 'Speculations on Perceptrons and Other Automata'. Presented at the RC-115, 2 June 1959. https://gwern.net/doc/ai/nn/1959-good.pdf. Moreover, it should be noted that the "intelligence explosion" account is one of three long-standing approaches to characterizing the features and impacts of the "singularity": other accounts include "an 'accelerating change' school, associated with Kurzweil, [and] an 'event horizon' school, associated with Vinge"; Chalmers, David J. 'The Singularity: A Philosophical Analysis'. *Journal of Consciousness Studies* 17 (2010): 7–65. Ftn 5; referring to: Yudkowsky, Eliezer. 'Three Major Singularity Schools'. Machine Intelligence Research Institute, 30 September 2007. https://intelligence.org/2007/09/30/three-major-singularity-schools/. See also (on the "accelerating change" account): Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin Books, 2006.; and (on the "event horizon" account): Vinge, Vernor. 'The Coming Technological Singularity: How to Survive in the Post-Human Era', 1993. https://edoras.sdsu.edu/~vinge/misc/singularity.html.

[329] Chalmers, David J. 'The Singularity: A Philosophical Analysis'. *Journal of Consciousness Studies* 17 (2010): pg. 7. http://consc.net/papers/singularityjcs.pdf

[330] Muehlhauser, Luke, and Anna Salamon. 'Intelligence Explosion: Evidence and Import'. In *Singularity Hypotheses*, edited by Amnon H. Eden, James H. Moor, Johnny H. Søraker, and Eric Steinhart, 15–42. The Frontiers Collection. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. https://doi.org/10.1007/978-3-642-32560-1_2. See also Muelhauser, Luke. 'Intelligence Explosion FAQ'. Machine Intelligence Research Institute, 2013. https://intelligence.org/ie-faq/.

[331] Anthropic. 'Anthropic's Responsible Scaling Policy, Version 1.0', 19 September 2023. Pg. 14.

[332] Ibid.

[333] Karnofsky, Holden. 'The Duplicator'. Cold Takes, 20 July 2021. https://www.cold-takes.com/the-duplicator/.

Systems with critical hazardous capabilities

Systems that pose or enable critical levels of (extreme or even existential) risk,[334] regardless of whether they demonstrate a full range of human-level/like cognitive abilities.

→ **Advanced AI**:

→ "systems substantially more capable (and dangerous) than existing [...] systems, without necessarily invoking specific generality capabilities or otherwise as implied by concepts such as 'Artificial General Intelligence.'"[335]

→ "Systems that are highly capable and general purpose."[336]

→ **High-risk AI system":**

→ An AI system that is both "(a) … intended to be used as a safety component of a product, or is itself a product covered by the Union harmonisation legislation [...] (b) the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or putting into service of that product [...]."[337]

→ "AI systems that are used to control the operation of critical infrastructure… [in particular] highly capable systems, increasingly autonomous systems, and systems that cross the digital-physical divide."[338]

[334] For a definition of "extreme" risks in this context, along with a review of some potentially dangerous specific capabilities of general-purpose models, see also: 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. Pg. 3 ("We focus on 'extreme' risks, i.e. those that would be extremely large in scale (even relative to the scale of deployment). This can be operationalised in terms of the scale of impact (e.g. damage in the tens of thousands of lives lost, hundreds of billions of dollars of economic or environmental damage) or the level of adverse disruption to the social and political order. The latter could mean, for example, the outbreak of inter-state war, a significant erosion in the quality of public discourse, or the widespread disempowerment of publics, governments, and other human-led organisations").

[335] 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/. Ftn 5.

[336] 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. Ftn 1.

[337] European Commission. 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts'. European Commission, 21 April 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Article 6(1)(a-b). Note, Article 6(2) further specifies "in addition to the high-risk AI systems referred to in paragraph 1, AI systems referred to in Annex III shall also be considered high-risk." Note, recent amendments adopted by the European Parliament introduce small changes to this definition: see European Parliament. 'Amendments Adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))'. European Parliament, 14 June 2023. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html. Note however that this classification of AI systems as "high-risk" is less focused on those systems' capabilities, and rather on their application areas and use cases. For a discussion of the overall implications of the EU AI Act's construction of "high-risk AI systems", see also: 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.

[338] Microsoft. 'Governing AI: A Blueprint for the Future', 2023. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw , pg. 14 (further clarifying that "one place to start might be to focus on AI systems that:

→ **AI systems of concern**: "highly capable AI systems that are [...] high in 'Property X' [defined as] intrinsic characteristics such as agent-like behavior, strategic awareness, and long-range planning."[339]

→ **Prepotent AI**: "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*."[340]

→ **APS Systems**: AI systems with "(a) Advanced capabilities, (b) agentic Planning, and (c) Strategic awareness."[341] These systems may risk instantiating "MAPS"—"misaligned, advanced, planning, strategically aware" systems;[342] also called "power-seeking AI".[343]

→ **WIDGET:** "Wildly Intelligent Device for Generalized Expertise and Technical Skills."[344]

→ "**Rogue AI**:

→ Take decisions or actions affecting large-scale networked systems;
→ Process or direct physical inputs and outputs;
→ Operate autonomously or semi-autonomously; and
→ Pose a significant potential risk of large-scale harm, including physical, economic, or environmental harm.").

[339] Matteucci, Kayla, Shahar Avin, Fazl Barez, and Seán Ó hÉigeartaigh. 'AI Systems of Concern'. arXiv, 9 October 2023. https://doi.org/10.48550/arXiv.2310.05876. pg. 1. (Also noting that: "We believe there is sufficient commonality amongst the cluster of *instrumental rationality*, *agency*, the mix of *agentic planning and strategic awareness*, and similar properties such as *consequentialism*, to mark it as Property X, which is strongly linked to potential intrinsic danger from advanced AI systems, such as the *pursuit of convergent instrumental goals* and the emergence of *power seeking behaviour*. Very likely this is not a single property, but rather a cluster of linked characteristics, which may evolve in time", Pg. 3). (emphasis in original).

[340] Critch, Andrew, and David Krueger. 'AI Research Considerations for Human Existential Safety (ARCHES)', 29 May 2020. http://acritch.com/arches/. pg 12–13. (emphasis in original).

[341] Carlsmith, Joseph. 'Is Power-Seeking AI an Existential Risk?' arXiv, April 2021. http://arxiv.org/abs/2206.13353. Pg. 8. (defining these terms as:

→ "*Advanced capability*: they outperform the best humans on some set of tasks which when performed at advanced levels grant significant power in today's world (tasks like scientific research, business/military/political strategy, engineering, and persuasion/manipulation).
→ *Agentic planning*: they make and execute plans, in pursuit of objectives, on the basis of models of the world.
→ *Strategic awareness*: the models they use in making plans represent with reasonable accuracy the causal upshot of gaining and maintaining power over humans and the real-world environment.").

[342] Leung, Jade. 'Priorities in AGI Governance Research'. Presented at EA Global: SF 22, 30 July 2022. https://www.listennotes.com/podcasts/ea-radio/priorities-in-agi-governance-WM_DUyzNPqR/.

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

[344] Chan, Alan. 'A Prosaic Case for Not Building AGI - Part I'. Substack newsletter. *Alan's Substack* (blog), 20 January 2023. https://coordination.substack.com/p/a-prosaic-case-for-not-building-agi. (referring to a system that has the following properties:

→ "The ability to plan over long time horizons, on the order of months to years.
→ Human-level or above fluency with language and language-based tasks, including coding.
→ Competence in interacting directly with the world digitally, rather than having interactions mediated through humans. This competence includes a general understanding of how the world works, but may not include skills like physical manipulation of objects.").

  - → "an autonomous AI system that could behave in ways that would be catastrophically harmful to a large fraction of humans, potentially endangering our societies and even our species or the biosphere."[345]

  - → "a powerful and dangerous AI [that] attempts to execute harmful goals, irrespective of whether the outcomes are intended by humans."[346]

- → **Runaway AI:** "advanced AI systems that far exceed human capabilities in many key domains, including persuasion and manipulation; military and political strategy; software development and hacking; and development of new technologies [...] [these] superhuman AI systems might be designed to autonomously pursue goals in the real world."[347]

- → "**Frontier (AI) model** *[relative-capabilities threshold]*:

  - → "large-scale machine-learning models that exceed the capabilities currently present in the most advanced existing models, and can perform a wide variety of tasks."[348]

  - → "highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models."[349]

- → **Frontier (AI) model** *[unexpected-capabilities threshold]*:

  - → "Highly capable foundation models, which could have dangerous capabilities that are sufficient to severely threaten public safety and global security. Examples of capabilities that would meet this standard include designing chemical weapons, exploiting vulnerabilities in safety-critical software systems, synthesising persuasive disinformation at scale, or evading human control."[350]

  - → "models that are both (a) close to, or exceeding, the average capabilities of the most capable existing models, and (b) different from other models, either in terms of scale, design (e.g. different architectures or alignment techniques), or their resulting mix of capabilities and behaviours."[351]

[345] Bengio, Yoshua. 'How Rogue AIs May Arise'. *Yoshua Bengio* (blog), 23 May 2023. https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/.

[346] Bengio, Yoshua. 'AI and Catastrophic Risk'. *Journal of Democracy* 34, no. 4 (2023): 111–21. https://muse.jhu.edu/pub/1/article/907692 pg. 113-114.

[347] Davidson, Tom. 'The Danger of Runaway AI'. *Journal of Democracy* 34, no. 4 (2023): 132–40. https://muse.jhu.edu/pub/1/article/907694 pg. 133.

[348] 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/.

[349] UK Government. 'AI Safety Summit: Introduction'. GOV.UK, 25 September 2023. https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html.

[350] 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. Pg. 6. Notably, they caution that this proposed definition "is lacking in sufficient precision to be used for regulatory purposes and that more work is required to fully assess the advantages and limitations of different approaches. Further, it is not our role to determine exactly what should fall within the scope of the regulatory proposals outlined – this will require more analysis and input from a wider range of actors." ibid. Pg 9.

[351] 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. Pg. 3. (in particular noting that such systems might generate dangerous capabilities, such as cyber-offense, deception, persuasion and

→ **Highly-capable systems of concern**:

→ "Highly capable foundation models [...] capable of exhibiting dangerous capabilities with the potential to cause significant physical and societal-scale harm"[352]

Appendix 2: Lists of definitions for policy tools and field

2.A. Terms for tools for intervention

Strategy[353]

→ **AI strategy research**: "the study of how humanity can best navigate the transition to a world with advanced AI systems (especially transformative AI), including political, economic, military, governance, and ethical dimensions."[354]

→ **AI strategy**: "the study of big picture AI policy questions, such as whether we should want AI to be narrowly or widely distributed and which research problems ought to be prioritized."[355]

→ **Long-term impact strategies**: "shape the processes that will eventually lead to strong AI systems, and steer them in a safer direction."[356]

→ **Strategy**: "the activity or project of doing research to inform interventions to achieve a particular goal. [...] AI strategy is strategy from the perspective that AI is important, focused on interventions to make AI go better."[357]

→ **AI macrostrategy:** "the study of high level questions having to do with prioritizing the use of resources on the current margin in order to achieve good AI outcomes."[358]

manipulation, political strategy, weapons acquisition, long-range horizon planning, AI development, situational awareness, self-proliferation, and others).

[352] 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. Ft. 6.

[353] The term was more characteristic of earlier work in the field, but is also still used somewhat widely. For instance, DeepMind operates a "Long-term Strategy & Governance" team. Vishal. 'DeepMind Is Hiring Long-Term Strategy & Governance Researchers'. EA Forum, 13 September 2021. https://forum.effectivealtruism.org/posts/atbonGDAFegfeDbTF/deepmind-is-hiring-long-term-strategy-and-governance.

[354] Flynn, Carrick. 'Personal Thoughts on Careers in AI Policy and Strategy'. Effective Altruism Forum, 27 September 2017. https://web.archive.org/web/20210622160148/https://forum.effectivealtruism.org/posts/RCvetzfDnBNFX7pLH/personal-thoughts-on-careers-in-ai-policy-and-strategy.

[355] Brundage, Miles. 'Guide to Working in AI Policy and Strategy'. 80,000 Hours, 13 June 2017. https://80000hours.org/articles/ai-policy-guide/.

[356] Hendrycks, Dan, and Mantas Mazeika. 'X-Risk Analysis for AI Research'. arXiv, 21 July 2022. http://arxiv.org/abs/2206.05862. Pg. 6.

[357] Stein-Perlman, Zach. 'Framing AI Strategy'. AI Impacts, 6 February 2023. https://aiimpacts.org/framing-ai-strategy/.

[358] See informally: Gabs, Nick. 'We Need Holistic AI Macrostrategy'. EA Forum, 15 January 2023. https://www.lesswrong.com/posts/Jh8Trhc89JDPjnk2J/we-need-holistic-ai-macrostrategy.

Policy

- → **AI policy**: "concrete soft or hard governance measures which may take a range of forms such as principles, codes of conduct, standards, innovation and economic policy or legislative approaches, along with underlying research agendas, to shape AI in a responsible, ethical and robust manner."[359]

- → **AI policymaking strategy**: "A research field that analyzes the policymaking process and draws implications for policy design, advocacy, organizational strategy, and AI governance as a whole."[360]

Governance

- → **AI governance**:

  - → "AI governance (or the governance of artificial intelligence) is the study of norms, policies, and institutions that can help humanity navigate the transition to a world with advanced artificial intelligence. This includes a broad range of subjects, from global coordination around regulating AI development to providing incentives for corporations to be more cautious in their AI research."[361]

  - → "local and global norms, policies, laws, processes, politics, and institutions (not just governments) that will affect social outcomes from the development and deployment of AI systems."[362]

  - → "shifting and setting up incentive structures for actions to be taken to achieve a desired outcome [around AI]."[363]

  - → "identifying and enforcing norms for AI developers and AI systems themselves to follow. [...] AI governance, as an area of human discourse, is engaged with the problem of aligning the development and deployment of AI technologies with broadly agreeable human values."[364]

  - → "the study or practice of local and global governance systems—including norms, policies, laws, processes, and institutions—govern or should govern AI research, development, deployment, and use."[365]

[359] 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.

[360] 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. Pg 3.

[361] EA Forum. 'AI Governance'. Accessed 21 November 2022. https://forum.effectivealtruism.org/topics/ai-governance.

[362] Muelhauser, Luke. 'Our AI Governance Grantmaking so Far'. Open Philanthropy, 16 December 2020. https://www.openphilanthropy.org/blog/ai-governance-grantmaking. See also 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.

[363] Leung, Jade. 'Priorities in AGI Governance Research'. Presented at the EA Global: SF 22, 30 July 2022. https://www.listennotes.com/podcasts/ea-radio/priorities-in-agi-governance-WM_DUyzNPqR/.

[364] 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#AI_governance__definition_

[365] 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-

→ **Collaborative governance of AI technology**: "collaboration between stakeholders specifically in the legal governance of AI technology. The stakeholders could include representatives of governments, companies, or other established groups."[366]

→ **AGI safety and governance practices**: "internal policies, processes, and organizational structures at AGI labs intended to reduce risk."[367]

2.B. Terms for the field of practice

AI governance

→ "the field of AI governance studies how humanity can best navigate the transition to advanced AI systems, focusing on the political, economic, military, governance, and ethical dimensions."[368]

→ "AI governance concerns how humanity can best navigate the transition to a world with advanced AI systems. It relates to how decisions are made about AI, and what institutions and arrangements would help those decisions to be made well."[369]

→ "AI governance refers (1) descriptively to the policies, norms, laws, and institutions that shape how AI is built and deployed, and (2) normatively to the aspiration that these promote good decisions (effective, safe, inclusive, legitimate, adaptive). [...] governance consists of much more than acts of governments, also including behaviors, norms, and institutions emerging from all segments of society. In one formulation, the field of AI governance studies how humanity can best navigate the transition to advanced AI systems."[370]

Transformative AI governance

→ "[governance that] includes both long-term AI and any nearer-term forms of AI that could affect the long-term future [and likewise] includes governance activities in both the near-term and the long-term that could affect the long-term future."[371]

and-cooperative-ai-development ftn 3 (citing ÓhÉigeartaigh, Seán S., Jess Whittlestone, Yang Liu, Yi Zeng, and Zhe Liu. 'Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance'. *Philosophy & Technology*, 15 May 2020. https://doi.org/10.1007/s13347-020-00402-x.).

[366] Critch, Andrew, and David Krueger. 'AI Research Considerations for Human Existential Safety (ARCHES)', 29 May 2020. http://acritch.com/arches/. Pg. 81.

[367] 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. Pg. 3.

[368] 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 5.

[369] Dafoe, Allan. 'AI Governance: Opportunity and Theory of Impact', 17 September 2020. https://www.allandafoe.com/opportunity.

[370] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. 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/9780197579329.013.2. Pg 1.

[371] 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.

Longterm(ist) AI governance

→ **Long-term AI governance**: "[governance that] includes both long-term AI and any nearer-term forms of AI that could affect the long-term future [and likewise] includes governance activities in both the near-term and the long-term that could affect the long-term future."[372]

→ **Longtermist AI governance**:

→ "longtermism-motivated AI governance / strategy / policy research, practice, advocacy, and talent-building."[373]

→ "the subset of [AI governance] work that is motivated by a concern for the very long-term impacts of AI. This overlaps significantly with work aiming to govern transformative AI (TAI)."[374]

→ "longtermist AI governance [...] which is intellectually and sociologically related to longtermism [...] explicitly prioritizes attention to considerations central to the long-term trajectory for humanity, and thus often to extreme risks (as well as extreme opportunities)."[375]

[372] 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.

[373] 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. Ftn 3. (crediting the term to Allan Dafoe).

[374] 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.

[375] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. 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/9780197579329.013.2.

Appendix 3: Auxiliary definitions and terms

Beyond this, it is also useful to clarify a range of auxiliary definitions that can support analysis in the advanced AI governance field. These include, but are not limited to:[376]

→ **Strategic parameters**: Features of the world that significantly determine the strategic nature of the advanced AI governance challenge. These parameters serve as highly decision-relevant or even crucial considerations, determining which interventions or solutions are appropriate, necessary, viable, or beneficial to addressing the advanced AI governance challenge; accordingly, different views of these underlying strategic parameters constitute underlying cruxes for different theories of actions and approaches. This encompasses different types of parameters:

→ **technical parameters** (e.g., advanced AI development timelines and trajectories, threat models, and feasibility of alignment solution),

→ **deployment parameters** (e.g., the distribution and constitution of actors developing advanced AI systems), and

→ **governance parameters** (e.g., the relative efficacy and viability of different governance instruments).[377]

→ **Key actor**: An actor 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).

→ **Key decision**: 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. This can include direct decisions about deployment or testing during a critical moment, but also includes many upstream decisions (such as over whether to initiate risky capabilities).

→ **Lever (of governance)**:[378] 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, and (3) other *key actors*' choices or *key decisions.*

→ **Pathway (to influence)**: A tool or intervention by which other actors (that are not themselves key actors) can affect, persuade, induce, incentivize, or require *key actors* to make certain *key decisions*. This can include interventions that ensure that certain *levers of control* are (not) used, or used in particular ways.

→ **(Decision-relevant) asset**: Resources that can be used by other actors in pursuing *pathways of influence* to *key actors*, and that aim to induce how these key actors make *key decisions* (e.g., about

[376] This list is not exhaustive, and there are of course many other relevant terms and concepts that feature in different governance approaches. See for instance Pistillo, Matteo. 'Compute Governance Key Concepts' (forthcoming draft).

[377] See also the model of different risk factors in: 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.

[378] See also Maas, Matthijs: 'Advanced AI Governance: A Literature Review'. *Institute for Law & AI*, Foundations Report 4. November 2023.

whether or how to use their *levers*). This includes new technical research insights, worked-out policy *products*; networks of direct influence, memes, or narratives;

→ **(Policy) product**: A subclass of *assets*; specific legible proposals that can be presented to *key actors*.

→ **Critical moment(s)**: High-leverage[379] moments where high-impact decisions are made by some actors on the basis of the available *decision-relevant assets*, which affect whether *beneficial advanced AI outcomes are within reach*. These critical moments may occur during a public "AI crunch time,"[380] but they may also occur potentially long in advance (if they lock in choices or trajectories).

→ **"Beneficial" AI outcomes**: The desired and/or non-catastrophic societal outcomes from AI technology. This is a complex normative question, which one may aim to derive by some external moral standard or philosophy,[381] through social choice theory,[382] or through some legitimate (e.g., democratic) process by key stakeholders themselves.[383] However, this concept is often undertheorized and needs significantly more work, scholarship, and normative and public deliberation.

[379] Another term that has been proposed for this is "hingey": 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.

[380] For a working definition of "crunch time", see: 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. ("a period lasting 1–20 years when the decisions most impactful on TAI outcomes might be made"); see also Tyre, Eli. 'How Do We Prepare for Final Crunch Time?' LessWrong 2.0, 30 March 2021. https://www.lesswrong.com/posts/wyYubb3eC5FS365nk/how-do-we-prepare-for-final-crunch-time. For additional discussion and distinguishing of different versions, see also Hadshar, Rose. 'What's Going on with "Crunch Time"?' EA Forum, 20 January 2023. https://forum.effectivealtruism.org/posts/7CdtdieiijWXWhiZB/what-s-going-on-with-crunch-time.

[381] See for example 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.

[382] Baum, Seth D. 'Social Choice Ethics in Artificial Intelligence'. *AI & Society* 35, no. 1 (March 2020): 165–76. https://doi.org/10.1007/s00146-017-0760-1.

[383] For one discussion of procedural requirements for politically legitimate governance for AI, see also 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.

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://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://law-ai.org/ai-policy-metaphors

→ Maas, Matthijs, “Advanced AI governance: A literature review.” *Institute for Law & AI*, **AI Foundations Report 4**. (November 2023). https://law-ai.org/advanced-ai-gov-litrev

INSTITUTE
FOR LAW & AI

law-ai.org