caisa Det Nationale Center for AI i Samfundet
The National Center for AI in Society

Artificial General Intelligence and its Societal Implications

Authors
Søgaard, Anders; Rogers, Anna;
Adler-Nissen, Rebecca; Belongie, Serge;
Moeslund, Thomas B.; Jurowetzki, Roman.

Editors
Søgaard, Anders; Feldt, Johannes N.

- Publication date: 11 November 2025
- Published by: The National Center for Artificial Intelligence in Society (CAISA), Denmark
- Document version: Publisher's PDF, also known as version of record
- Citation for published version (APA): Søgaard, A., Rogers, A., Adler-Nissen, R., Belongie, S., Moeslund, T., & Jurowetzki, R. (2025). *Artificial General Intelligence and its Societal Implications*. The National Center for Artificial Intelligence in Society (CAISA). https://caisa.dk/forskning

Artificial General Intelligence and its Societal Implications

By Anders Søgaard, Anna Rogers, Rebecca Adler-Nissen, Serge Belongie, Thomas B. Moeslund and Roman Jurowetzki

Summary

The prospect of artificial general intelligence (AGI) has led some researchers and entrepreneurs to foresee an imminent "intelligence explosion," leading to either humanity's near-extinction, or a future of abundance. This brief examines the philosophical and technical premises underpinning such predictions, and their implications for governance. There are two such premises: first, that AI systems can be intelligent and become more intelligent with time; and second, that advanced AI can direct its behavior towards autonomously set goals. However, these premises rest on relatively weak foundations. CAISA recommends a clear-headed approach to AGI, as long as the concept and governance suggestions remain vague. The rush to prioritize AGI development—and/or its governance—risks diverting attention from the tangible risks posed by existing AI technologies, such as algorithmic bias, privacy erosion, and unchecked power concentrations. At the same time, to have a more meaningful discussion of AGI further research is needed, spanning philosophy, cognitive science, cybersecurity as well as the societal and political impact of AGI narratives.

Background

Currently, the main technology people refer to as 'artificial intelligence' (AI) is generative language models. Language models were *not* originally presented as 'intelligent', but their success on tasks like question answering (especially on tests designed to assess human knowledge, like mathematical olympiads) is why they are labeled 'AI' in public discourse today. The term 'AI' was historically motivated by marketing rather than scientific goals but has taken on a life of its own. The same development has now led many researchers and entrepreneurs to speculate that we are at the verge of a technological revolution, associated with a so-called intelligence explosion and the arrival of 'artificial general intelligence' (AGI). These terms remain vague, but such speculations take up a significant fraction of day-to-day news and global policy-making. In 2023, open letters signed by leading tech executives and researchers warned that AGI would pose "profound risks" and even "risk of human extinction." Researchers and entrepreneurs who expect such a revolution have predicted dramatic phase transitions for societies, from *near-extinction* to *abundance*. Many of these predictions recycle ideas from science fiction and earlier discussions of technology, but in this brief, we focus on the perils of modern society having to navigate responsible governance in the midst of such predictions. While the predictions seem outlandish to many researchers, many in power take them seriously, and that means we should seriously consider what these predictions turn on. Who presents us with these predictions and why? What assumptions lie behind such predictions? Importantly, the worries raised by these predictions draw attention away from the immediate challenges around disinformation, privacy, job displacement, de-skilling, copyright, climate, etc. - a topic we will return to.

Mapping the philosophical landscape

We take the central premises behind the predictions of near-extinction or abundance to be:

1. AI systems can be intelligent and become more intelligent with time (leading to an **intelligence explosion**).
2. AI systems can direct its behavior towards **autonomously set goals**.

What supports the two premises? This we turn to next. We use the two premises to induce a four-way taxonomy of positions around AI and safety.

Adherence to Premise 1 and the idea of an intelligence explosion implies **realism about intelligence**, i.e., the view that intelligence is not just a catch-all gesture to a bundle of skills of social status, but refers to a real process or capacity with causal impact. Intelligence realism is common in AI [1], yet controversial elsewhere [2,3]. It is also common to think that intelligence is substrate-independent, i.e. can exist in non-biological systems, but this is not implied by intelligence realism. We distinguish between realists and non-realists about artificial intelligence, but there are two groups of non-realists, intelligence realists who believe intelligence is substrate-dependent, and non-realists about intelligence. Intelligence realists who believe intelligence is substrate-independent, often also believe in a forthcoming intelligence explosion, since AI systems train and infer faster than humans – and therefore arguably improve faster over time. Superintelligence refers to the idea that over time AI models will surpass human-level intelligence (with no ceiling in sight). Superintelligence thus implies realism about intelligence, as well as no upper bounds on whatever intelligence refers to.

What about Premise 2? Language models and related technologies are designed to follow instructions. Many instrumental goals may serve such overall goals, but do such goals emerge on their own, and to what extent will models consistently follow them? Do models exhibit consistent, goal-directed behavior? Broadly speaking, AI researchers have approached this question by either observing model behavior or probing their internals. So far, both approaches run into a series of methodological problems [4,5], e.g., how to sample from the set of relevant scenarios, how to define goals at a fixed level of granularity, and how to provide criteria for membership. Most evidence for goal-directed model behavior has been anecdotal in nature. In terms of stances, we again have two positions: those who believe in consistent goal-directed behavior (i.e. that current models or close-by extensions thereof can meaningfully develop autonomy and for instance plot against humans) – and those who do not, or who do not believe we are in a position to have reasons to believe one or the other.

We summarize the possible positions with respect to both premises as follows (see table 1).

Table 1: AI and safety, four positions.

	AI can set autonomous goals	*AI cannot set autonomous goals*
Intelligence is real, and AI can have it	**doomers:** models will become superintelligent and may plot against us	**bloomers:** models will become superintelligent but cannot plot against us
Intelligence is not real, or AI cannot have it	**deemers:** models will not become superintelligent but may plot against us	**bleemers:** models will not become superintelligent and cannot plot against us[1]

Note that even if Premise 1 (intelligence explosion) and Premise 2 (emergent goals) both hold, models still need venues of direct influence. This is of course subject to material limitations. For example, a risk scenario may involve a drone equipped with an on-device model. A material limitation is in this case introduced by the computing machinery that we can fit onto a drone. Another risk scenario may involve a model copying itself across servers, beyond the control of its developer. A physical limitation would then be induced by the

[1] Note that the 'bleemer' position does not necessarily entail the view that models are harmless: a major disruption could happen without either 'plotting' or 'intelligence', e.g. via errors in generated code controlling physical infrastructure, or via user manipulation scenarios observed in training data. One area of particular concern is *agents*: models increasingly deployed to act autonomously in the real world, e.g. by generating and executing code without human oversight. Even if the harmful behaviors stem purely from patterns in the training data, and the models have no real 'intelligence', there could still be significant negative impact.

bandwidth available for connecting the two, and their capacity. Currently, a medium-sized language model with 70 billion parameters takes up about 130GB of space when stored in 16-bit precision. To use it, it needs to fit in memory. This requires significant computer infrastructure, preventing models from copying themselves onto smaller devices for now.

The political landscape

While Premise 1 and 2 have limited scientific support [2,6], doomers and bloomers already influence policy-makers [7], e.g. some countries have given up creator protections to avoid stymying progress, and in the US, the construction of AI data processing centers has been used explicitly as a reason to increase fossil fuel-based energy production [8]. The race narrative implies that developing it before our adversaries is more important than estimating its impact. While China has distanced itself from the AGI discourse to focus on being best at adoption, many US-based tech companies argue that the "AGI geopolitical race" requires unchecked development, and lobby against the regulation of current technology and use.

The risk induced by the hypothetical AGI comes in many flavors, but all risks turn on what venues the models can directly influence, and the physical limitations of the infrastructure available to them. The kind of risks of concern to doomers depend on an intelligence explosion (Premise 1) and emergent goals (Premise 2). To respond to the relevant risk scenarios, we can thus intervene not only on Premise 1 and Premise 2, but also on direct venues of influence and physical limitations. We should therefore also try to understand what venues and infrastructure limitations are in place, and what cybersecurity measures can/should be developed to keep models in their intended scope. Some of these questions remain surprisingly unexplored.

Others [9] have introduced a four-way taxonomy of governance strategies (i–iv): Either we count on a (i) technical or (ii) cultural plateau slowing down progress, or we control the models through (iii) alignment or (iv) oversight. For example, most safety research falls in the third bucket, whereas EU regulation has focused on the fourth. The plateaus challenge Premise 1, while 'alignment' and oversight are responses to Premise 2. Research on alignment and oversight is already well-funded, but more research is needed to evaluate whether technical or cultural plateaus are on the horizon. The fifth governance strategy would be a global moratorium.

Regulation through alignment, oversight, or a moratorium introduces several challenges, including defining what counts as AGI, getting adversaries to trust that regulation is effective, that all states observe the rules, and that violations are punished. Proponents of alignment tend to argue from both Premise 1 and 2, whereas proponents of oversight tend to only subscribe to Premise 1, with slow take-off. Suggestions for governing the hypothetical AGI should be evaluated in relation to tangible benefits and risks of current forms of AI. In particular, the following should be considered.

- *Overlooking current risks*: the push to govern "existential" AGI risks often comes from actors who also benefit from minimal regulation of today's AI technologies. We need to be aware that those voices advocating for future-focused governance often simultaneously resist stricter oversight of current AI systems even though they display tangible harms like bias, privacy violations, job displacement, adverse impacts on education and mental health, misinformation spread, and increased cybersecurity threats.

- *Opportunity costs*: billions are currently invested in AGI research and "alignment" (making AGI safe). This detracts from efforts to tailor current models for applications such as optimizing energy use, fighting climate change or improving healthcare. More research is also needed to establish the optimal strategies for human oversight, AI-assisted workflows, and accountability across sectors.

Main points and implications

Most safety work around AI buys into Premise 1 and Premise 2, but relatively little research has been devoted to evaluating these premises, as well as understanding the material or physical limitations of AI, what we suggest calling Premise 3. Based on the above, we foreground four main points and their implications. Our first point is not to let concerns around AGI overshadow concerns around the negative impact of current AI technologies. Our other three points concern the lack of knowledge and research surrounding: the cognitive science and philosophy of intelligence and goal-directedness; the physical limitations of current and future models and their potential impact on society and its infrastructures; and how AGI narratives already influence local and global politics.

- Since predictions around AGI turn on two premises for which support is weak, we recommend a sober approach. This does not mean a slow-down of research or roll-out, but we should not sacrifice or

compromise regulation and democratic oversight of existing AI technology in fear of losing the 'AGI race'.

- To enable a more meaningful discussion of AGI, we need to better understand the philosophy and cognitive science of intelligence and goal-directedness. What would it take for a system—whether biological or digital—to instantiate these capabilities?

- We need better analysis of the physical limitations of the current and next generation models, and the ways in which they could directly impact society and its infrastructures. This calls for more support for education and research in cybersecurity, as well as collaboration between cybersecurity, mechanistic interpretability, and cognitive science.

- We need better analysis of how different AGI narratives and expectancies already shape political decisions, locally and globally. This calls for dedicated funding for independent research in the public interest, on such topics as global dynamics of funding AI research and deployment, the influence of AGI narratives on researchers, policy- and lawmakers, and how this influence is enacted.

AGI has been center stage in many recent public debates. This discussion raises important research questions, and talk of AGI may itself have geopolitical consequences. At the same time, the debates have taken airtime from discussions of other societal implications of AI roll-out, including copyright, privacy, higher education, and mental health. CAISA recommends a clear-headed approach to AGI, as long as the very concept of AGI and the suggestions for how to govern it remain vague.

About the authors

Anders Søgaard is Professor of Artificial Intelligence and Machine Learning at the Department of Computer Science, Head of Center for Philosophy of Artificial Intelligence at the University of Copenhagen and chief scientist at CAISA. Anna Rogers is Associate Professor of Data Science at the IT University of Copenhagen and chief scientist at CAISA. Rebecca Adler-Nissen is Professor of International Relations at the Department of Political Science, University of Copenhagen and Director of CAISA. Serge Belongie is Professor of Computer Science at the Department of Computer Science, Director of The Pioneer Centre for Artificial Intelligence, University of Copenhagen, and chief scientist at CAISA. Thomas B. Moeslund is Professor of Media Technology at the Department of Architecture, Design, and Media Technology at Aalborg University and Deputy Director of CAISA. Roman Jurowetzki is Associate Professor of Innovation and Applied Data Science at Aalborg University Business School and chief scientist at CAISA.

About CAISA

The National Center for Artificial Intelligence in Society (CAISA) is a national consortium that gathers researchers from the University of Copenhagen, Aalborg University, Aarhus University, the IT University of Copenhagen and the Technical University of Denmark in close collaboration with the Pioneer Centre for Artificial Intelligence (P1).

As Denmark's independent research center for artificial intelligence in society, CAISA centers the citizen. We carry out groundbreaking interdisciplinary research, and we deliver overview of the most recent scientific breakthroughs. Based on new and interdisciplinary research, we advise decision-makers in the public and private sectors on how they may develop and use artificial intelligence practically, such that it contributes to growth, supports democracy and strengthens digital autonomy.

CAISA is funded for a three-year pilot period with 20 million Danish kroner from the Danish Ministry of Digital Affairs, and 30 million Danish kroner from the national research reserve.

References

1. François Chollet. “On the Measure of Intelligence”. arXiv:1911.01547. Preprint, arXiv, 25. november 2019.
2. Godfrey H. Thomson “A Hierarchy Without a General Factor.” *British Journal of Psychology, 1904-1920* 8, no. 3 (1916): 271–81.
3. Timnit Gebru, and Émile P. Torres. “The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence.” *First Monday*, ahead of print, April 14, 2024.
4. Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. “We Need to Talk About Random Splits.” *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Association for Computational Linguistics, 2021, 1823–32.
5. Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” *Nature Machine Intelligence* 1 no. 5 (2019): 206–215.
6. Nina Rajcic and Anders Søgaard, “Goal-Directedness Is in the Eye of the Beholder,” arXiv:2508.13247, preprint, arXiv, August 18, 2025.
7. Seán Ó hÉigeartaigh. “The Most Dangerous Fiction: The Rhetoric and Reality of the AI Race”. SSRN Scholarly Paper No. 5278644. Social Science Research Network, 25. maj 2025.
8. Donald. J. Trump, “Reinvigorating America’s Beautiful Clean Coal Industry and Amending Executive Order 14241,” The White House, April 8, 2025, https://www.whitehouse.gov/presidential-actions/2025/04/reinvigorating-americas-beautiful-clean-coal-industry-and-amending-executive-order-14241/.
9. Herman Cappelen, Simon Goldstein, and John Hawthorne. “AI Survival Stories: A Taxonomic Analysis of AI Existential Risk”. *Philosophy of AI* 1 (June 2025): 1–19.