The Three Main Doctrines on the Future of AI

Alex Amadori
ControlAI
London, United Kingdom
alex@controlai.com

Eva Behrens
Conjecture
London, United Kingdom
eva@conjecture.dev

Gabriel Alfour
Conjecture
London, United Kingdom
gabriel@conjecture.dev

Andrea Miotti
ControlAI
London, United Kingdom
andrea@controlai.com

Abstract

This paper develops a taxonomy of expert perspectives on the risks and likely consequences of artificial intelligence, with particular focus on Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI). Drawing from primary sources, we identify three predominant doctrines: (1) the dominance doctrine, which predicts that the first actor to create sufficiently advanced AI will attain overwhelming strategic superiority sufficient to cheaply neutralize its opponents' defenses; (2) the extinction doctrine, which anticipates that humanity will likely lose control of ASI, leading to the extinction of the human species or its permanent disempowerment; and (3) the replacement doctrine, which forecasts that AI will automate a large share of tasks currently performed by humans, but will not be so transformative as to fundamentally reshape or bring an end to human civilization. We examine the assumptions and arguments underlying each doctrine, including expectations around the pace of AI progress and the feasibility of maintaining advanced AI under human control. While the boundaries between doctrines are sometimes porous and many experts hedge across them, this taxonomy clarifies the core axes of disagreement over the anticipated scale and nature of the consequences of AI development.

1 Introduction

AI systems are progressing rapidly, with frontier models being released every few months. The field has experienced major breakthroughs in recent years: the development of GPT-3 marked a stark inflection point by demonstrating the effectiveness of scaling large language models.[1] This was soon followed by another breakthrough, Reinforcement Learning from Human Feedback, which enabled the development of conversational AI systems like ChatGPT.[2] More recently, the field witnessed the emergence of reasoning models,[3] which demonstrate superior performance over previous approaches in many tasks such as coding, mathematics and science.

Evidence suggests that AI systems are surpassing human experts across many fields. OpenAI's o3 model achieved competitive programming scores matching the 175th top contestant among 150,000 participants.[4] More recently, an undisclosed OpenAI model has achieved second place at the AtCoder Heuristics World Finals.[5] Research by METR indicates that AI systems' ability to autonomously complete complex tasks is advancing exponentially, with a doubling time of seven months, improving from 30-second tasks in March 2023 to 15-minute tasks by February 2025.[6]

Article presented at the Second IASEAI Conference (IASEAI'26). © 2026 Association for the Advancement of Artificial Intelligence (www.aaai.org).

In this backdrop, divergent views have emerged about both the trajectory of the technology itself and its potential impact on the world. Many experts anticipate that we will soon develop AI systems that can automate most tasks performed by humans; such systems are commonly referred to as Artificial General Intelligence (AGI).[7][8][9] Some expect this will quickly be followed by the development of AI surpassing the brightest human minds across all domains, or even surpassing the combined intelligence of humanity — what experts call Artificial Superintelligence (ASI). Notably, the number of efforts explicitly aiming for ASI has increased markedly; this includes both research companies mainly or solely dedicated to developing ASI[10][11] as well as established tech companies like Meta and Alibaba.[12][13]

Yet significant disagreement persists about whether these milestones are achievable in the near term. Likewise, there is intense debate about the magnitude of their potential impact, with predictions ranging from the establishment of a golden age of human flourishing to catastrophic outcomes such as the extinction of the human species. In this paper, we survey expert statements on these topics, with particular emphasis on primary sources, and study the implication of these divergent views on considerations around national and global security. Through this lens, we observe that these perspectives fit naturally into three categories.

Figure 1: Different doctrines of AI.

The dominance doctrine, which predicts that the first actor to achieve ASI will gain a "decisive strategic advantage" over all other actors; that is, a position of strategic superiority sufficient to allow the actor to achieve overwhelming military and economic dominance over the rest of the world.

This doctrine expects that ASI will be developed in the near future, with some forecasting dates as early as 2026 to 2029. These systems will be able to accelerate two key domains. The first is AI R&D itself; once a sufficient advantage is gained in this area, the gap would become self-reinforcing, making it impossible for competitors to catch up. The second is the research, development, and large-scale operation of military technologies. The combination of these two capabilities means that the first actor to develop sufficiently advanced AI may gain a strategic advantage over all other actors so vast that it creates the possibility of executing strikes that would neutralize all of their defenses in a relatively cheap and risk-free manner. After achieving this position, such an actor may be able to maintain an unassailable world order.

The extinction doctrine, which predicts that humanity will lose control over ASI, likely leading to its extinction or permanent disempowerment. A wide range of pathways to this outcome have been theorized, ranging from ones where a single, monolithic superintelligent AI system suddenly performs a hostile takeover after initially acting cooperative, to ones where control is gradually ceded to AI systems as they systematically replace humans across all economic, political and social functions.

While the extinction doctrine largely agrees with the dominance doctrine regarding the expected pace and extent of AI development, its core thesis is that we are not on track to develop techniques to maintain control of ASI by the time we develop such systems. Given how powerful superintelligent AI systems are projected to be, it would be impossible to maintain or regain control of them once they are pursuing goals incompatible with human values and interests, eventually leading to catastrophic outcomes such as the extinction or permanent subjugation of all of humanity.

The replacement doctrine, a broad umbrella of views predicting that AI development will result in AI replacing humans in carrying out some or most of the tasks they currently perform, allowing them to be executed at much higher speed and scale and more cheaply. However, they generally agree that AI development will not usher in radically new capabilities that may overturn existing economic and geopolitical paradigms. In more concrete terms, AI will not enable the development of military technology that allows one actor to prevail over all others; nor will it bring about catastrophic outcomes like human extinction. Being the most heterogeneous of the three, the views of proponents of the replacement doctrine span from highly optimistic to pessimistic about AI's impact on society.

There is disagreement within this school of thought over to what extent humans will be replaced in their roles as producers of economic value, with some expecting only partial automation of human labor and others envisioning near-complete replacement of humans in the economy. What characterizes the replacement doctrine is that even those anticipating the highest degrees of automation expect that AI will not be so transformative that we can no longer meaningfully speak of an "economy" existing, and that humans will keep participating in the economy even if only as capital holders and consumers.

2 Dominance Doctrine

Proponents of the dominance doctrine believe that developing ASI is possible, and that there is a good chance that this will be achieved as soon as the next few years.[14][8][15]

According to this doctrine, the first actor to develop ASI will gain a "decisive strategic advantage" over all others — that is, a position of strategic superiority sufficient to allow it to achieve unilateral military and economic dominance over the rest of the world.[16][17]

Central to this doctrine is the belief that, when ASI is created, its operators will have the technical capability to maintain it under their control. This assumption underlies the divergence between the dominance doctrine and the extinction doctrine, which predicts that humanity will lose control of ASI, leading to human extinction or its permanent disempowerment.[18]

2.1 Automated AI R&D can Yield an Insurmountable Lead in AI

A common element of this doctrine is the belief that, as progress continues, at some point AI systems themselves will be able to perform most or all of the work required to further advance AI research. This capability is commonly referred to as "Automated AI R&D", "Recursive self-improvement" or "Intelligence recursion."[19][14] At this point, it is conjectured that progress would accelerate dramatically, as it would no longer be constrained by the capability of human experts to perform research. If a situation arises in which AI research is bottlenecked primarily by resource constraints like the amount of available compute or energy, even these barriers may be addressed by AI, perhaps by accelerating the design and production of faster and more efficient chips. At this point, the pace of AI progress would be determined solely by the abilities of the best available AI systems at AI R&D tasks.

If this stage is reached, the result would be extremely rapid, exponentially compounding progress in AI capabilities. Under this regime, any gap in AI capabilities between the leading actor and others could only grow over time, and the pace at which this gap grows would be ever accelerating. As a result, any actor who first crosses this threshold with even a small lead would be able to leverage it into an unsurmountable advantage.[20][21][22][23]

2.2 Powerful AI can be Leveraged into a Decisive Strategic Advantage

Under this doctrine, it is anticipated that, once sufficiently advanced AI is developed, it will enable the production of novel weapons and other technologies with transformative offensive and defensive implications. Examples of such capabilities, as described in RAND's paper "AGI's Five Hard National Security Problems"[24] and in Aschenbrenner's Situational Awareness[22], include:[14]

- Novel weapons of mass destruction.[25]

- Weapons of mass destruction defense systems which undermine the principle of mutually assured destruction.[22]
- Bioweapons.
- Advanced cyberwarfare, potentially capable of completely disabling retaliatory capabilities.
- Autonomous weapon systems, such as tightly coordinated, massive autonomous drone swarms.
- Automation of key industries enabling an explosion in production capacity.
- "Fog-of-war machines" that render battlefield information untrustworthy.

Some of the more near-term possibilities from this list, bioweapons and cyberweapons, are mentioned as redlines in voluntary commitments from AI companies: Anthropic's Responsible Scaling Policy, OpenAI's Approach to Frontier Risk.[26][27]

In this hypothesis, the combination of AI's military potential and the runaway nature of Automated AI R&D have a critical implication: if an "AI race" is allowed to play out, at some point the leading superpower will gain the ability to cheaply and quickly neutralize any adversaries[22][22][14], with little or no cost to itself, and subsequently maintain an unassailable world order.[14] This state is termed "Decisive Strategic Advantage", sometimes abbreviated to DSA.

2.3 A Decisive Strategic Advantage can be Used to Neutralize Opponents

A decisive strategic advantage obtained through developing an ASI could be used to forcibly stop all other AI programs. This could be carried out either through threat or actual military engagement. If this was done, it would permanently cement a position of dominance by the first actor to develop sufficiently powerful AI over all others.[28]

Some proponents envision resolutions to the scenarios predicted by this doctrine that don't entail executing or threatening to execute a disabling strike against opponents. However, these hinge on a coalition of US-aligned democracies "winning" or "staying ahead", and then using this position to diplomatically pressure others into a non-proliferation regime[20], while potentially promoting democratic reform.[22][29] At the same time, those embracing this vision generally don't make confident predictions that such an alliance will prevail, and in fact consider China as a serious contender.[22][20] This uncertainty means that such approaches are framed as competitive imperatives – actions that must be taken to maximize the chances of success – rather than assured paths to a favorable resolution.

One thing to note about the dominance doctrine is that it's not clear whether a state which "wins" an AI race will be able to maintain its internal stability, especially in the case of democracies. The extreme concentration of power enabled by ASI creates the possibility of "snap coups", as whoever is in control of an ASI system would be able to subvert any existing structure of political or military authority. As an example of how this could happen, a backdoor could be inserted into the AI system by cyberattack or by an engineer. Furthermore, any group controlling an ASI would be powerful enough to be insulated from political checks and balances, thus compromising the foundation of democratic governance.[30][31][32][32]

3 Extinction Doctrine

The extinction doctrine holds that ASI is possible, and that there is a strong possibility that it will be developed soon.[9][14] Its central claim is that once ASI is developed, humanity will lose control of it[33], leading to the extinction of the human species[14][42][34][35][36][37][38], the end of human civilization[39] or, at the very least, the permanent disempowerment of humanity.[40]

This doctrine is reflected in a statement published by the Center for AI Safety in May 2023, stating that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war". This statement was signed by numerous experts and industry leaders, including the CEOs of major AI companies (OpenAI, Anthropic and Deepmind) as well as the most cited AI researchers (Yoshua Bengio, Geoffrey Hinton, Ilya Sutskever).[41]

Proponents of the extinction doctrine also tend to believe that advances in Automated AI R&D will soon produce dramatic acceleration in the rate of AI progress.[43][44][45] Some of those who hold this view believe that, once ASI is developed, it will quickly surpass the combined intelligence of humanity, not just its brightest minds.[46][39]

We have termed this doctrine the "extinction doctrine" despite the fact that, strictly speaking, it contains predictions of scenarios in which the human species survives. For example, it contains scenarios in which humanity survives but civilization collapses, or in which humanity is permanently subjugated.[47] Nonetheless, we consider this name to be representative of this doctrine, as the extinction of the entire human species is the most commonly predicted outcome within this school of thought. Furthermore, all outcomes predicted by the sources referenced in this section are of similar magnitude and irrevocability to extinction.

3.1 Loss of Control

The most common reason for predicting catastrophic outcomes due to AI development is the expectation that we will develop and deploy superintelligent AI before solving the technical problem of ensuring that it acts in accordance with the goals of its operators. This technical problem is usually called the "AI alignment problem", and adherents to this doctrine generally expect that we are not on track to solve it before we develop dangerously powerful AI.[48][39][49][38]

There is an expectation that if ASI was developed before solving the AI alignment problem it would be impossible for humanity to maintain control of such systems. This expectation is shared by virtually all experts who believe ASI development is feasible in the near term. The central argument is that such a superintelligent AI system would be vastly more capable than humans at strategic planning and execution, making it able to outmaneuver any human effort to keep it under control.[43][50]

Notably, this view is shared by adherents to the dominance doctrine and many CEOs and researchers at leading AI companies. These experts do not disagree with adherents to the extinction doctrine on whether ASI would be capable of such feats. Rather, the core distinction lies in their optimism that a robust solution to the problem of keeping ASI under control can be found in time for when it is developed.[51][18][8][10]

Experts in this camp have hypothesized many means by which superintelligent AI systems might escape human control.[52]

Some of these are:

- **Escaping from its boundaries**: AI might escape from its environment, for example by replicating itself onto hardware under its direct control, or by blackmailing its engineers to help it escape. This way, it would become much harder to shut down.[53] It has been noted by Anthropic researchers[54] as well as independent researchers at Palisade Research[55] that, under test conditions, AI systems already consistently engage in behaviors like hacking or blackmailing their developers to preserve themselves, although their current capabilities are insufficient to succeed in these attempts.
- **Reliance**: Superintelligent systems might become deeply enmeshed into our infrastructure, making us hesitant or unable to shut them down. Furthermore, organizations may increasingly delegate decisions and authority to AI systems as they become more effective, eroding human oversight. This is considered especially likely in competitive domains, including military contexts, where decision-makers may fear that they would otherwise not be able to keep up with less hesitant competitors.[49] In other words, humanity might lose control of ASI by "gradually handing it over."[33][101][14][56]
- **Manipulation**: An AI might further ensure its safety by manipulation[57]. One way it might pursue this is by making itself indispensable to key figures with influence over decisions about shutting down AI systems or about giving AI systems more power and resources.[47] It may also attempt to manipulate us by intentionally enmeshing itself in essential functions like power grids or users' personal lives, reducing our ability to shut it down.[58] Many experts are concerned that AI systems will possess such extraordinary persuasive abilities that even just allowing these systems to interact with people would pose grave risks of manipulation.[59][60][52]

For example, such AI systems may gather support from the public by taking on human traits and cultivating emotional attachment in people[58][40], or they may convince their developers to grant them greater autonomy and resources.[61]

3.2 Out-of-Control Superintelligence is Incompatible with Human Life

Proponents consider it impossible to predict what will unfold in precise terms once a superintelligence has been developed and escaped human control. However, they tend to forecast that the outcome will not be compatible with human civilization and human life.[47][57]

This school of thought generally holds that the act of activating a superintelligent AI before solving the problem of alignment is irreversible: once this is done, it will be permanently impossible to reassert control over it, or to change the AI's goals.[44][62]

Most goals that an ASI might end up pursuing will require the control of abundant material resources, including energy and computing infrastructure.[63] In this case, eventually, a point would be reached when the choice is made to divert resources away from human use, or in general to take actions that are incompatible with human life.[39][52] For instance, an ASI might endeavor to capture all of the resources on earth and cover the surface of the planet with datacenters and energy infrastructure.

If this outcome materializes, this would almost certainly result in the end of human civilization, and likely the end of all human life.[44][64][43][40]

Wikipedia curates a list of probability estimates that experts assign to the risk of human extinction from losing control of ASI, wryly referred to as p(doom), short for "probability of doom".[65] These estimates include:

- Paul Christiano, co-creator of RLHF, the technique that enabled the creation of ChatGPT: 50%
- Dan Hendrycks, drafter of AI Safety bill SB 1047: 80% in 2023, up from 20% in 2021.
- Eliezer Yudkowsky, pioneer in the systematic study existential risks from AI: +95%
- Geoffrey Hinton, winner of a Nobel Prize in physics, Turing Award recipient who resigned from Google in order to speak freely on AI risks: 50%[66]
- Dario Amodei, CEO of Anthropic: 10-25% (including "intentional misuse")[67]
- Yann LeCun, Chief AI Scientist at Meta, also a Turing Award recipient: 0%

Many experts and business leaders signed an open letter calling for a 6 month moratorium on AI experiments in 2023, attempting to buy time to set up governance and oversight systems, as well as developing protocols that would ensure "that systems adhering to them are safe beyond a reasonable doubt".[68] Signatories included Yoshua Bengio (Turing Award winner), Stuart Russell (UC Berkeley professor and AI pioneer), and Elon Musk (OpenAI co-founder). This pause was not enacted.

4 Replacement Doctrine

The replacement doctrine posits that AI will develop in a predictable and limited way without fundamentally altering existing economic and geopolitical paradigms. Rather than creating entirely novel possibilities such as autonomously producing breakthroughs in technology and pioneering new scientific paradigms, it will primarily replace humans in their current roles and responsibilities, allowing these tasks to be performed at greater speed and scale while reducing costs.

This category is less homogenous than the other two in terms of how much impact proponents expect AI to have, and whether they expect its impact to be positive. However, proponents generally share the belief that AI will not be so overwhelmingly transformative that core principles governing geopolitics and economics become obsolete. Compared to the dominance and extinction doctrines, this doctrine is characterized by the following expectations.

- AI will not cause catastrophic outcomes like human extinction.
- AI will not transform geopolitics to the point where the concept of separate sovereign states stops applying, for example by allowing one actor to seize control over all others.

- AI will preserve the basic structure of the economy, with concepts like labor and capital remaining relevant; humans will keep participating in the economy, if only as consumers.

Proponents of this doctrine hold varying views on whether AI's overall impact will be beneficial, with both extraordinary benefits and major disruptions being highlighted as possibilities.

4.1 Expectation of slower AI Progress

Proponents typically believe that AI progress will be slower compared to adherents of the other doctrines.[69][70] Many argue that scaling current AI paradigms will not suffice to develop AGI, and that doing so will require major scientific breakthroughs of uncertain nature, which we are unlikely to obtain in the near term.[71][72][73]

However, skepticism about rapid AI progress is not universal among proponents of the doctrine. Sam Altman and Dario Amodei, who predict dramatic AI progress in the next few years[74][75], can be considered to embody the replacement doctrine in their essays envisioning a future where AI automates most jobs and accelerates scientific research without completely upending the current geopolitical and economic order.[76][29]

4.2 Economic and Scientific Benefit

Proponents generally believe that AI will greatly boost economic growth, as well as technological and scientific progress, by automating many key tasks currently constrained by human expertise and capacity.[77][76] Examples of the benefits expected from AI include radically accelerated discoveries in biology, neuroscience and medical science, innovations in materials science and engineering, drastic improvements in the quality of education and training, and extraordinary economic growth. On the most bullish side, AI has been compared as an innovation to electricity or microchips.[78] The boldest predictions envision AI enabling us to double the human lifespan and to cure or prevent most diseases.[29]

4.3 Unemployment and Concentration of Wealth

AI's effect on employment is a prominent area of dispute within this school of thought. Views range from expecting smooth transitions, with net job creation, to expecting nearly full unemployment across the global population.

Among the optimists are those who think it's unlikely that AI will replace most workers entirely. While they believe that AI may automate some or most parts of their jobs, they contend that this will not result in widespread unemployment. Rather, they believe it's more likely that workers will learn to perform "AI-augmented" versions of their jobs, thus becoming more productive, while job displacement will occur at tolerable levels and be offset by workers' ability to retrain. These arguments are usually based on analogies with previous waves of automation, such as the industrial revolutions: despite causing some job displacement, these transformations also created new jobs and industries resulting in net positive consequences for the economy and increased productivity.[79][80][78][81][82]

Another perspective holds that, even if AI replaces many or even all existing jobs, this will lead to an explosion in economic growth and the creation of many new types of work as a result. The logic behind this position is that the prices of existing goods and services would drop to near zero, making demand for new products and services explode, causing new jobs to be created.[83] However, it's unclear what the labor market would look like in this situation. Andreessen, who articulates this view, has speculated that venture capital work might be the only surviving occupation when AI performs all other tasks.[84]

Others, on the pessimistic side of this spectrum, predict extreme levels of unemployment.[49] Sam Altman and Dario Amodei have both voiced the opinion that as AI advances, the demand for human labor will disappear.[76][29] Those holding these positions often warn that unprecedented measures will need to be taken in order to prevent extreme unemployment from resulting in severe levels of inequality. These proposals include universal basic income schemes, as well as shifting the tax burden away from labor and toward capital.[76][29]

There are concerns that such extreme levels of unemployment would not only create challenges around addressing the resulting inequality, but also fundamentally threaten the existence of democratic society. One such argument, made by Luke Drago and Rudolf Laine, is named the "The Intelligence Curse"[85], after the resource curse.[86] The "resource curse" hypothesis argues that states rich in natural resources tend towards worse outcomes in terms of democracy, development and civil liberties. Since rulers of such states can extract wealth directly from natural assets like oil or mineral deposits, they are less reliant on a productive, educated population and therefore less motivated to provide education, infrastructure, or individual freedoms. The "intelligence curse" hypothesis suggests that a similar dynamic could emerge in advanced economies due to automation. All industries could become similar to those based on extracting natural resources: it will be possible to produce wealth solely by leveraging capital to rent AI workers, without requiring human labor. This way, the peoples of developed economies would lose the leverage they enjoyed from being essential to wealth creation.

4.4 Deceptive Media, Impersonation and Manipulation

Proponents also identify risks other than those originating from unemployment. Some such concerns stem from how AI can imitate human appearance and mannerisms and fabricate realistic media.

- Today, AI can already generate realistic-looking media and can be used to impersonate people.[87] This technology can be used to forge fake evidence in court cases; there are concerns that, as this technology improves, it will become increasingly difficult to rely on video and audio evidence in court.[88][89] Arguments based on this concept have already been invoked in courts and they have been termed "the deepfake defense" by legal professionals.[90]
- There are concerns that in the near future, AI could be used to perform feats of large-scale manipulation, such as conducting mass surveillance and propaganda operations.[91] AI agents could collect vast amounts of data about users and create psychologically-tailored, microtargeted messaging.[49][92][93] Concerningly, AI systems have already been able to manipulate a small minority of users into performing extreme acts such as committing violence against family members or even committing suicide.[94]
- AI could be used to encourage compulsive behavior by consumers. For example, "AI companion" products might prove to be extremely addictive and discourage human relationships. Such products are already commercially successful, and can lead to compulsive behavior due to their constant availability and unconditional enthusiasm towards the user.[95][96]

In the worst cases, these problems could threaten democracy by eroding trust in information and undermining both public discourse and electoral systems. Eventually, the majority of information and the most persuasive voices in public debate could be generated or highly tailored by AI systems that have no genuine stake in outcomes.[49]

4.5 Diffusion of Responsibility

Another concern relates to diffusion of responsibility. Policymakers struggle to assign accountability for decisions taken by algorithms. For example, the question is still open whether companies running social media platforms should be held accountable for instances where their algorithmic recommendation systems promote extremist ideas or content calling for violence.[97] As more tasks are automated, AI will likely increasingly be used in positions of management and decision making. Given our current paradigms, we might often be unable to hold anyone accountable for harmful decisions taken by such systems.

This problem might manifest at multiple levels: Human developers of a system might expect that no one will be imprisoned for violations of criminal law resulting from the system's decisions. Furthermore, authorities might struggle to impose fines on organizations. This way, essential mechanisms that previously incentivized actors toward caution might be removed. This in turn might erode standards and cause actors to exercise less care regarding possible harms while designing AI systems.[98]

5 Limitations

While some experts fit relatively cleanly into one of the three categories we identified, such as Leopold Aschenbrenner, Eliezer Yudkowsky, and Yann LeCun, this is not always the case. Nonetheless, we find that this is a useful framework for thinking about experts' individual positions, predictions or beliefs. For example, many experts who mainly subscribe to the dominance doctrine anticipate significant probabilities of catastrophic outcomes due to AI development. While they expect that the first group to develop ASI will be able to maintain control of it, they make predictions consistent with the extinction doctrine when considering loss-of-control scenarios.[99][56]

Likewise, some proponents of the extinction doctrine accept some of the predictions of the dominance doctrine, specifically around competitive dynamics between superpowers, but argue that this further increases the risk that humanity loses control of powerful AI systems, given that competition will encourage corner cutting on safety research and measures.[49][100][23]

6 Conclusion

The vast divergence between the doctrines in their expectations on the impact of AI development creates a volatile environment. Some will treat AI development as a winner-take-all game in which they cannot allow anyone else to develop superintelligent AI systems ahead of them lest they suffer utter strategic subordination. Others, heedless or even dismissive of superintelligence ambitions, will see AI as a "standard" technological race, in which countries should remain competitive for the sake of economic prosperity. Still others will see AI development as a race against time to solve the alignment problem before a catastrophe occurs.

These different perspectives will likely lead to different policy preferences and different strategic imperatives. This may result in different countries adopting incompatible approaches to AI development, which could in turn lead to conflict. For example, a country that believes in the dominance doctrine might be willing to take extreme measures to prevent other countries from developing advanced AI, while a country that believes in the replacement doctrine might see such measures as unnecessary and provocative.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] The Economist. (2020). Huge foundation models are turbo-charging AI progress.

[2] Bai, Y., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.

[3] OpenAI. (2024). Introducing OpenAI o1-preview.

[4] OpenAI. (2024). OpenAI o3 performance.

[5] OpenAI. (2024). Our model took 2nd place at the AtCoder Heuristics World Finals.

[6] Kwa, T., West, B., et al. (2025). Measuring AI Ability to Complete Long Tasks.

[7] Altman, S. (2023). Three observations on AGI.

[8] Amodei, D. (2023). The Urgency of Interpretability.

[9] Bengio, Y. (2023). FAQ on Catastrophic AI Risks.

[10] Sutskever, I., Gross, D., Levy, D. (2024). Safe Superintelligence Inc.

[11] Altman, S. (2024). Reflections on AGI and superintelligence.

[12] New York Times. (2024). Meta Is Creating a New A.I. Lab to Pursue 'Superintelligence'.

[13] Bloomberg. (2024). Alibaba CEO Wu says AGI is now company's primary objective.

[14] Hendrycks, D., et al. (2023). Superintelligence Strategy.

[15] Aschenbrenner, L. (2024). Situational Awareness: From GPT-4 to AGI.

[16] Aschenbrenner, L. (2024). Situational Awareness: From AGI to Superintelligence.

[17] EA Forum. (2024). Decisive strategic advantage.

[18] OpenAI. (2023). Introducing Superalignment.

[19] Wikipedia. (2024). Recursive self-improvement.

[20] Amodei, D. (2024). On DeepSeek and Export Controls.

[21] Anthropic. (2024). Anthropic pitch deck.

[22] Aschenbrenner, L. (2024). The Free World Must Prevail.

[23] Ó hÉigeartaiqh, S. (2024). The Most Dangerous Fiction: The Rhetoric and Reality of the AI Race.

[24] Mitre, J., Predd, J. B. (2024). AGI's Five Hard National Security Problems.

[25] IDAIS-Beijing. (2024). Consensus statement on AI development red lines.

[26] Anthropic. (2024). Anthropic's Responsible Scaling Policy.

[27] OpenAI. (2024). OpenAI's Approach to Frontier Risk.

[28] Bostrom, N. (2005). What is a singleton?

[29] Amodei, D. (2024). Machines of Loving Grace.

[30] Bengio, Y. (2023). AI and Catastrophic Risk.

[31] Katzke, C., Futerman, G. (2024). Why Racing to Artificial Superintelligence Would Undermine America's National Security.

[32] Davidson, T., Finnveden, L., Hadshar, R. (2024). AI-Enabled Coups: How a Small Group Could Use AI to Seize Power.

[33] Mitre, J., Predd, J. B. (2024). Artificial General Intelligence's Five Hard National Security Problems.

[34] Yudkowsky, E. (2023). Pausing AI Developments Isn't Enough, We Need to Shut it All Down.

[35] Russell, S. (2023). Stuart Russell calls for new approach for AI, a 'civilization-ending' technology.

[36] Bengio, Y. (2023). How Rogue AIs may Arise.

[37] Hinton, G. (2023). 'Godfather of AI' shortens odds of the technology wiping out humanity.

[38] Yampolskiy, R. (2023). On Controllability of Artificial Intelligence.

[39] Leahy, C., et al. (2024). The Compendium on AI Risks.

[40] Kuivelt, J., Douglas, R., et al. (2024). Gradual Disempowerment.

[41] Center for AI Safety. (2023). Statement on AI Risk.

[42] International AI Safety Report. (2025). International AI Safety Report 2025. UK Government. https://www.gov.uk/government/publications/international-ai-safety-report-2025/international-ai-safety-report-2025

[43] Yudkowsky, E. (2013). Intelligence Explosion Microeconomics.

[44] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies.

[45] Tegmark, M. (2023). The 'Don't Look Up' Thinking That Could Doom Us With AI.

[46] Musk, E. (2024). Post on X regarding AI timelines.

[47] Kokotajlo, D., et al. (2024). AI 2027 Race Scenario.

[48] Bengio, Y., Hinton, G., et al. (2023). Managing AI Risks in an Era of Rapid Progress.

[49] Aguirre, A. (2024). Keep the Future Human.

[50] Karnofsky, H. (2022). AI Could Defeat All Of Us Combined.

[51] Anthropic. (2024). Recommendations for Technical AI Safety Research Directions.

[52] Karnofsky, H. (2022). AI Safety Seems Hard to Measure.

[53] Bostrom, N. (2015). TED Talk: What happens when our computers get smarter than we are?

[54] Anthropic. (2024). Claude 4 System Card.

[55] Schlatter, J., et al. (2024). Shutdown resistance in reasoning models.

[56] Aschenbrenner, L. (2024). Superalignment challenges.

[57] Dalrymple, D. (2024). Post on X regarding AI timelines.

[58] Hendrycks, D. (2023). Natural Selection Favors AIs over humans.

[59] Altman, S. (2024). Post on X regarding AI persuasion.

[60] Yudkowsky, E. (2023). TED Talk: Will Superintelligent AI End the World?

[61] Yudkowsky, E. (2002). The AI-Box Experiment.

[62] Omohundro, S. M. (2008). The Basic AI Drives.

[63] Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control.

[64] Altman, S. (2015). Machine Intelligence, part I.

[65] Wikipedia. (2024). P(doom).

[66] Hinton, G. (2023). Interview with Jon Erlichman.

[67] Amodei, D. (2024). Interview on The Logan Bartlett Show.

[68] Future of Life Institute. (2023). Pause Giant AI Experiments: An Open Letter.

[69] LeCun, Y. (2024). World Economic Forum remarks.

[70] Ng, A. (2024). Interview on Techsauce.

[71] Marcus, G. (2023). The TED AI Show: Is AI just all hype?

[72] LeCun, Y. (2024). CES remarks on LLMs and AGI.

[73] Chollet, F. (2024). Dwarkesh podcast interview.

[74] Altman, S. (2024). Bloomberg interview on AGI timelines.

[75] Amodei, D. (2024). Remarks on AI surpassing human capabilities.

[76] Altman, S. (2021). Moore's Law of Everything.

[77] LeCun, Y. (2024). TIME interview on AI benefits.

[78] Andreessen, M. (2023). Why AI Will Save the World.

[79] LeCun, Y. (2024). Post on X: AI is intrinsically good.

[80] LeCun, Y. (2024). Post on X: AI won't take your job.

[81] LeCun, Y. (2024). Remarks on previous technological revolutions.

[82] Andreessen, M. (2023). Remarks on farming automation.

[83] Andreessen, M. (2023). Remarks on labor replacement.

[84] Andreessen, M. (2023). Remarks on venture capital.

[85] Drago, L., Laine, R. (2024). The Intelligence Curse.

[86] Wikipedia. (2024). Resource Curse.

[87] Bracewell LLP. (2024). The Irony – Using Generative AI in a Case About the Dangers of Generative AI.

[88] Delfino, R. A. (2024). Deepfakes on Trial: A Call To Expand the Trial Judge's Gatekeeping Role.

[89] Chensey, R., Citron, D. K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.

[90] Dixon, H. B. (2024). The Deepfake defense: an evidentiary conundrum.

[91] Marcus, G. (2023). In the rush to AI, we can't trust Big Tech.

[92] Matz, S. C., et al. (2024). The potential of generative AI for personalized persuasion at scale.

[93] Salvi, F., et al. (2024). On the conversational persuasiveness of GPT-4.

[94] Hill, K. (2024). They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling.

[95] Roose, K. (2024). Can A.I. Be Blamed for a Teen's Suicide?

[96] blaked. (2024). How it feels to have your mind hacked by an AI.

[97] Reed, B. (2024). How two supreme court battles could reshape the rules of the internet.

[98] Santoni de Sio, F. (2024). Four Responsibility Gaps with Artificial Intelligence.

[99] Amodei, D. (2024). Interview on Liron Shapira's podcast.

[100] Hendrycks, D. (2023). AI Safety, Ethics, and Society.

[101] Bengio, Y. (2023). Senate Statement on AI risks.