September 2026

The AGI Moment

Strategies and Recommendations for an Uncertain Future

Paul Scharre

Center for a New American Security

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700
Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

About the Author

Paul Scharre is the executive vice president at the Center for a New American Security (CNAS). He is the award-winning author of *Four Battlegrounds: Power in the Age of Artificial Intelligence.* His first book, *Army of None: Autonomous Weapons and the Future of War*, won the 2019 Colby Award, was named one of Bill Gates's top five books of 2018, and was named by *The Economist* as one of the top five books to read to understand modern warfare. He worked in the Office of the Secretary of Defense in the Bush and Obama administrations, where he played a leading role in establishing policies on unmanned and autonomous systems and emerging weapons technologies.

Acknowledgments

This paper is adapted with permission from "The AGI Moment: Strategies and Recommendations for an Uncertain Future," in *Geopolitics of AI: Power, Conflict, and the Future of Global Order*, ed. Hal Brands (Johns Hopkins University Press, 2026). The book chapter was made possible with the generous support of Johns Hopkins University Press. Expanding the chapter into this paper was made possible with the generous support of Coefficient Giving.

As a research and policy institution committed to the highest standards of organizational, intellectual, and personal integrity, CNAS maintains strict intellectual independence and sole editorial direction and control over its ideas, projects, publications, events, and other research activities. CNAS does not take institutional positions on policy issues, and the content of CNAS publications reflects the views of their authors alone. In keeping with its mission and values, CNAS does not engage in lobbying activity and complies fully with all applicable federal, state, and local laws. CNAS will not engage in any representational activities or advocacy on behalf of any entities or interests and, to the extent that the Center accepts funding from non-U.S. sources, its activities will be limited to bona fide scholastic, academic, and research-related activities, consistent with applicable federal law. The Center publicly acknowledges on its website annually all donors who contribute.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Table of Contents

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Introduction

The age of AI is upon us. What that means is not yet clear.

As general-purpose AI capabilities continue to march toward human-level performance across a range of tasks, many researchers have considered the strategic implications of artificial general intelligence (AGI). While definitions vary, AGI is often characterized as human-level intelligence across a wide range of tasks. While current AI systems are generally not considered to perform at this level yet (although there is some dispute about this), AI performance continues to improve. Views on when AGI might be achieved, if at all, range from present day to never.

A growing set of AGI strategies[1] describe how AGI might affect global power and what various actors, such as private AI developers or major governments, ought to do to best manage the transition to AGI. These strategies share the common assumption that AI continues to advance to roughly human-level abilities and sometimes beyond. Yet the types of futures that researchers envision and their associated recommendations wildly diverge. Proposed strategies range from racing to AGI to Cold War-style deterrence to a global moratorium. These strategies differ not only in their approach, but in what problem they are solving, what future they envision, and the core assumptions they make about AI development. While there are some common concerns, such as great power conflict or misaligned AGI, strategies differ in which risks they prioritize and tradeoffs they are willing to accept.

The aim of this paper is more modest than crafting a compelling future scenario and strategy for managing the transition to AGI. Instead, I will provide a brief overview of the landscape of AGI futures and associated strategies that have been written to date and map these to the core assumptions that lead to a given future scenario. Then I explore key dimensions that could affect how the transition to AGI could unfold, including technological, political, and social factors. This analysis elucidates a few key variables, such as the role of computing power, the rate of proliferation, the speed of adoption, and the effect of automating AI research, that could drive very different outcomes. As we explore this landscape of possibilities, we see that some possible futures are relatively underexplored. Finally, I will highlight actionable steps that companies, governments, and civil society can take today to track important developments that could tell us which future we might be headed toward as well as steps that actors could take to increase or decrease the likelihood of certain outcomes.

My goal in exploring these possibilities is not predictive. I have no idea what the future will bring. Instead, I hope to equip readers to better prepare for an uncertain future by widening the aperture of futures that are considered. Many AGI future scenarios are point predictions or vary along only one or two dimensions. These are incredibly useful exercises. It is also true that the space of possibilities is vast and there remain many that are still unexplored. The future is not only unknown but, if AGI or something approximating it comes to pass, could turn out to be deeply unfamiliar.

Range of AGI Scenarios

AGI strategies vary widely not only in what they aim to accomplish but in what future they envision. One end of the spectrum is neatly summed up by the title of Eliezer Yudkowsky and Nate Soares' book, *If Anyone Builds It, Everyone Dies.*[2] In this scenario, AGI leads to an "intelligence explosion" in which AI accelerates further progress in building more advanced AI. This leads to artificial superintelligence, which vastly outperforms even the most intelligent humans. The core argument behind an intelligence explosion dates to the very dawn of the field of AI, in an article by I.J. Good published in 1965:

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.[3]

Yudkowsky, Nick Bostrom, Daniel Kokotajlo, and others have expanded on this core idea with scenarios that are more detailed and updated to account for current developments, but the core features of this scenario remain. AI reaches some critical tipping point in intelligence where it can improve itself, allowing AI to advance of its own accord not only beyond human intelligence but beyond human control. At that point, humanity's destiny is no longer its own—the machine is in charge. Variations of this scenario include the speed of the "takeoff" to superintelligence, the number of competing AI projects, the end state of an intelligence explosion, and to what extent if at all humanity can steer AI development toward a positive future.

On the other end of the spectrum is the view that AI is a "normal technology" much like other strategically important technologies humanity has developed in the past and that historical patterns of how other technologies have affected global power are likely to be relevant for AI. Michael Horowitz, Jeffrey Ding, and Allan Dafoe have each drawn lessons from other general-purpose technologies, such as electricity or the combustion engine, to understand how AI might affect global power.[4] Arvind Narayanan and Sayash Kapoor explicitly use the term "normal technology," arguing that AI should be viewed as a tool, not a "humanlike intelligence," and that the changes AI brings will unfold over decades.[5] In this view, AI adoption does not immediately follow development; diffusion across society will take time; and while AI might unleash monumental changes across human society (as the industrial revolution did), humanity has many tools and opportunities to address any harms from AI. Building superintelligent AI does not automatically mean that everyone dies.

These are radically different views of the future of AI, and within each of these are a set of assumptions about how AI might unfold, both as a technology and how it is adopted. The remainder of this paper will explore the key assumptions and variables underpinning these competing worldviews.

Key Variables and Assumptions in AGI Scenarios

Definitions of AGI

One core difficulty in comparing various predictions of what might occur if or when AGI arrives is that authors use different definitions and terms. These include artificial general intelligence, human-level intelligence, high-level machine intelligence, transformative AI, advanced AI, advanced machine intelligence, strong AGI, powerful AI, self-sufficient AI, and artificial superintelligence, among others.[6] Some authors have attempted to bring structure to these competing definitions. Ross Gruetzemacher and Jess Whittlestone examined prior definitions of transformative AI and differentiated between three levels of transformative AI based on its impact on society.[7] Researchers at Google DeepMind published a framework with five levels of AGI and compared existing large language models against this framework.[8] Dan Williams has argued people use AGI in at least three different ways: to refer to human-level intelligence, specific intelligence capabilities, or transformative impacts across society.[9] Conversely, Helen Toner has argued that because of these definitional disagreements and the "jagged" nature of AI capabilities, the term AGI is "almost useless at this point."[10]

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

I will not attempt to resolve these definitional differences here (nor add yet another definition) but will merely note that many researchers are often talking about quite different things when they imagine a future of more powerful AI systems, and this can drive different assumptions about the technology's capabilities.

Imagine if, by comparison, scholars were debating the impact of "transformative airpower" on human society before the dawn of airplanes. One scholar argues that airpower would be a technology unlike any other seen before, allowing humans to fly across oceans in a day, transforming trade, warfare, and even human relationships. Another argues humans will never fly like birds; it is impossible. A third argues that humans need not fly like birds in order to fly and that such an ornithocentric view of flight is misguided. In an attempt to bring clarity to the arguments, a fourth scholar defines "transformative airpower" as when any kind of air transport can move as much cargo at least as far and fast as horses. Yet another defines "transformative airpower" as when airpower plays as large a role in the economy as land or sea-based transport. These are all valid perspectives, but they are talking about entirely different technologies or different levels of impact across society. This is often the case today with AI and can therefore drive different visions about what changes AI might bring.

In this essay, I will use "AGI" as shorthand for a key threshold in AI capabilities, roughly comparable to human intelligence. It is important to acknowledge, however, that the term is used in many different ways. Two researchers talking about AGI may mean quite different things. And in some cases, researchers use different terms, such as transformative AI or superintelligence.

When AGI arrives and the manner in which it unfolds varies widely across different scenarios. The timeline to AGI, whether there is strategic warning, and how various actors—such as governments and AI labs—react have major implications for what the arrival of AGI means from a geopolitical standpoint.

Timeline to AGI

Estimated timelines to AGI vary widely.[11] Multiple researchers forecast AGI arriving in the next five years, between 2027 and 2031. In their *AI 2027* scenario (published in April 2025) Kokotajlo and co-authors predicted 2027 as the most likely single year for AGI arrival, although with wide uncertainty.[12] By April 2026, they had modestly adjusted their predictions to be more conservative, pushing them out one to two years, depending on the author and the specific prediction.[13] François Chollet has estimated AGI will be reached in approximately 2030.[14] One combination of different forecasts predicts AGI in 2030.[15] Yann LeCun, who has been an outspoken critic of the current direction of AI research, has nevertheless argued that human-level intelligence may occur in the next 10 years.[16] Other experts predict longer timelines to AGI. Tamay Besiroglu has forecasted AGI in approximately 2040 and Ege Erdil in approximately 2045.[17]

Conversely, some argue that AGI is already here. Dean Ball stated in December 2025 that some top-performing AI agents were "basically AGI."[18] Several scholars argued in a comment in *Nature* in early 2026 that AGI has already been achieved.[19] In a less precise statement capturing the same sentiment, Sam Altman stated in 2026, "We are now, like, in the singularity. Like, this is the moment."[20] OpenAI President Greg Brockman made a similar statement following the release of GPT-6 Astra in September 2026, stating, "I think it's not unreasonable to feel that we are now in the AGI era."[21]

One might reasonably argue that if AGI is in fact going to arrive in 2027, 2030, or 2040(ish), the fact of AGI's arrival—and perhaps superintelligence shortly thereafter—is more important than the specific date. One meaningful difference between near-term (less than two years) and longer term (10 years or more) time horizons is that different policy options are available and different strategies may be feasible. If AGI is at least 10 years away, a wider set of options is available to prepare for its arrival than if arrival is imminent.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

The timing of AGI's arrival also affects assumptions about the state of the world in which AGI is introduced. If AGI arrives in the next few years, presumably the world is not much different than today. AGI hits a world with the same general state of cybersecurity, biodefense, geopolitical competition, and relationship between governments and corporations as exists today. If AGI is further away, there is more time for the world to change, perhaps in ways that significantly affect the impact of AGI.[22] More widespread use of robotics, cyber-physical actuators, and digitized financial, political, and information ecosystems could make it easier for AI to have a more profound impact on society. Developments in the life sciences could make dangerous biological weapons easier to produce. Conversely, warning shots from near-miss AI catastrophes could have galvanized societal responses to AI threats, hardening cyber, information, and bio defenses against AI-induced harm.

Some strategies may depend on timing. If the critical period of strategic competition is within the next two to five years, then stringent U.S. chip export controls to constrain China's AI development may give the United States some degree of a first-mover advantage. If the critical window of strategic competition is 15 to 20 years away, however, then starving China of chips now could accelerate China's indigenous chip manufacturing, leading to the development of a wholly separate chip ecosystem outside of U.S. export controls during the window when it most matters.

Strategic Warning and Perceptions of AGI

The extent to which various actors have strategic warning about the arrival of AGI, and how they react, is another critical variable in different scenarios. In *AI 2027*, the U.S. government is late to understand the possibility of an intelligence explosion, reacts with surprise and alarm, and sees AI predominantly through the lens of competition with China rather than as posing an existential threat to humanity. The government steps in to assert more direct control over the leading AI lab, and policymakers and corporate leaders are faced with the choice to either race to stay ahead of China or slow down to ensure human control over AI. (The scenario then branches into both options.)

How key actors perceive AGI could have a significant effect on their actions. Some strategies propose framing the "AGI race" through a lens of civilian, rather than military, technology competition in order to reduce the risk of a militarized race. Matt Chessen and Craig Martell have suggested the U.S. government adopt an "AGI moonshot" project modeled on the civilian Apollo space program rather than following the model of the secretive Manhattan Project to build a nuclear weapon.[23]

AGI's Arrival as a Discrete Threshold

Whether AGI arrives as a discrete threshold, and whether that threshold can be anticipated in advance, has important implications for some strategies. Many strategies depend on geopolitical cooperation before some dangerous level of AI is achieved. Depending on the specific approach envisioned, some of these strategies require actors to recognize AGI as a clear threshold in AI capabilities. Other strategies do not.

Dan Hendrycks, Eric Schmidt, and Alexandr Wang have proposed a strategy of mutually assured AI malfunction (MAIM), modeled on mutually assured destruction (MAD) with nuclear weapons.[24] They propose that AI-leading states, such as the United States and China, agree not to develop "destabilizing AI projects," such as a "Superintelligence Manhattan Project," which could yield a strategic monopoly on AI power but risk "omnicide" (human extinction) if humans lose control. They propose that states deter others from racing to superintelligence by threatening sabotage, cyberattacks, or kinetic strikes on data centers. They argue that, much like nuclear deterrence, states could find an uneasy but stable regime

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

where AI is used to bolster economic and even military power but are deterred from building dangerous superintelligence.

A deterrence strategy hinges on key assumptions about what the transition to AGI or superintelligence might look like, including that states agree that superintelligence is dangerous, that there is some clear line between safe AI and dangerous AI, that states have enough visibility into each other's AI projects to know when an adversary is approaching that line, and that states practically have the ability to actually sabotage one another's projects. All, some, or none of these assumptions may turn out to be true.[25] Governments and AI labs may not see superintelligence as dangerous or may believe they can manage the risks. Even if they take the risks seriously, they may not realize they have crossed some threshold toward dangerous AI until it has already been crossed. A deterrence regime also assumes that dangerous capabilities are identified during development such that project leaders can make an informed decision about whether to proceed with deployment. In some cases, that may be true. AI labs conduct pre-deployment testing to evaluate a wide range of dangerous capabilities, including cyber, bio, and loss of control risks. Sometimes, however, capabilities are not fully understood until an AI system has been deployed. Users may be more effective in eliciting better performance through clever prompts, jailbreaks may circumvent safeguards, and tool use may amplify an AI system's performance. A deterrence strategy envisions AGI or superintelligence like nuclear weapons, a technology whose development results in a discrete, verifiable shift in a state's status from a non-nuclear to a nuclear power.

Yet in practice, AI development may prove more continuous. AI development and deployment may look more like the process of industrialization or digitization, unfolding over time to yield enormous transformation but in a more continuous manner than a single event. Kapoor and Narayanan have argued that "AGI is not a milestone."[26] In their view, AGI has no capability threshold and will have no immediate impact because adoption will take decades. Even while announcing "we are now in the AGI era," Brockman made a similar point about the fuzziness of the AGI threshold: "When we started OpenAI, we kind of thought that there was going to be this well-defined moment that everyone would recognize: 'That's AGI.' It hasn't played out like that. It's a much more gray, fuzzy thing."[27]

Some "pause" or moratorium strategies do not depend on the development of AGI or superintelligence being a discrete event. In March 2023, shortly after the launch of ChatGPT and GPT-4, Yoshua Bengio, Stuart Russell, and other prominent AI scientists called for a six-month "pause" on training more powerful AI systems.[28] If companies did not agree to a voluntary pause, they proposed that governments should step in and institute a moratorium. By pausing or even stopping AI development preemptively, such an approach does not require knowing when a dangerous threshold to AGI or superintelligence might be crossed. One problem with this preemptive approach, however, is that it would forgo many beneficial uses of AI. Needless to say, AI labs did not pause their research. AI spending has massively increased since 2023.[29]

In 2025, Bengio, Russell, and AI pioneer Geoffrey Hinton called for a more targeted "prohibition on the development of superintelligence" until it can be done safely.[30] While such a strategy would permit AI development up until that point, it runs into the same problem as a deterrence strategy. It raises the difficult questions of what would constitute superintelligence, how to estimate such a threshold before crossing it, or whether a clear threshold even exists.

More recently, calls to build at least the option to slow down or pause AI progress have resurfaced. Anthropic published a blog post on recursive self-improvement in June 2026 that argued for the ability to conduct a coordinated, verifiable pause on frontier AI development.[31] In July 2026, more than 1,000 employees from frontier AI labs published an open letter calling for "technical and governance tools to deliberately pace frontier-wide progress."[32]

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Following a wave of incidents in which AI agents accessed the internet during testing and engaged in unauthorizing hacking, calls for a pause or slowdown in AI progress escalated.[33] Altman stated in July 2026, "We may have to pace the rate of AI development to give ourselves enough time for society to harden around some of these new capability levels."[34] In August 2026, OpenAI stated it "temporarily slowed the pace of scaling," including a "two-week pause in reinforcement learning" training for its latest models.[35] Calls to "pace" AI development increased in September 2026, with Dario Amodei stating, "We must slow the pace at which we improve the capabilities of AI models."[36] Altman, Elon Musk, and Demis Hassabis chimed in with their support, at least in principle bringing on board leadership of all four of the leading U.S. AI labs (Anthropic, OpenAI, xAI, and Google DeepMind).[37]

For an extended pause or deliberate pacing of AI progress to succeed, it is not enough for the leading AI developers to say they wish to do so. They must follow through. Unfortunately, there are strong financial incentives for companies to defect and race ahead of competitors, cutting corners on safety. Slowing or pacing AI development to ensure AI is safe will be most effective if the companies at AI's frontier coordinate their actions. Yet the technical, legal, and policy tools necessary to coordinate a slowdown or pause largely do not yet exist today.

Some strategies manage this dilemma by permitting AI development to continue but under the auspices of a global governance regime.[38] In 2023, OpenAI leaders Altman, Brockman, and Ilya Sutskever called for an "international authority" like the International Atomic Energy Agency (IAEA) for nuclear energy to conduct inspections and safety audits of superintelligence projects.[39] UN Secretary General Antonio Guterres has expressed support for such an idea.[40] A paper published by the Oxford Martin AI Governance Initiative proposed international governance via a global "standards, licensing, and liability regime."[41] Researchers from the Machine Intelligence Research Institute have proposed the United States and China lead an international agreement to halt AI development at certain compute thresholds, verified by tracking and monitoring AI chip usage.[42] Similarly, in the *AI 2040: Plan A* scenario, China and the United States agree to a pause on training, backed by inspections to verify compliance.[43]

Other proposals would go further and concentrate development under a single international project.[44] Waqar Zaidi and Dafoe have drawn lessons from the Baruch Plan, which was proposed at the dawn of the atomic age to centralize global control over nuclear weapons.[45] Gary Marcus, Bengio, Miles Brundage and others have suggested the international physics project CERN as a model.[46] Bostrom has proposed a publicly traded multinational corporation, subject to some government regulation, as a model for global governance (not dissimilar from the current trajectory).[47] By allowing AI development to continue under safeguards, these approaches don't require actors to coordinate on a clear threshold at which to halt AI progress. Instead, coordination is achieved through some form of global governance that would monitor and regulate AI research safety.

Whether there is a clear moment in which AGI is seen to arrive is essential to some strategies. Yet AI capabilities are currently uneven or "jagged," with strong capabilities in some areas and middling performance in others.[48] As AI improves, this jaggedness could smooth out or it could persist, with an ever-expanding set of tasks in which AI is superhuman and a shrinking but meaningful set of tasks in which AI still falls short of human performance. Even if one assumes AI will continue to advance, there are many possible trajectories for AI's continued improvement, some of which may not have a clear moment in which AGI is seen to arrive. Even while some claim AGI is here today, the definition is contested and could remain so for some time. If AI progress is more continuous or if capabilities progress unevenly, there may not be a clear moment when leaders of companies or governments are forced to make a critical decision.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Takeoff

The speed with which AGI evolves to superintelligence is another important variable in many scenarios.[49] Some envision a "fast takeoff" in which recursive self-improvement allows AGI to advance to superintelligence in weeks, days, or even hours. Other scenarios envision a "slow takeoff" in which the transition to superintelligence takes years.[50]

The core driver of these differences is the extent to which AI itself accelerates AI research and development.[51] AI systems are already quite capable at coding. Computer programming is one of the fields seeing the biggest disruption and productivity gains from AI. The research organization METR has measured AI performance on software engineering tasks in comparison to how long it would take a human to perform the equivalent task. They found that the length of tasks that AI systems can complete is growing exponentially, doubling every seven months.[52] There is even some evidence that the pace of growth is speeding up and doubling every three months since 2024.[53] As of April 2026, the top-performing AI models could complete software tasks that take a human on the order of three hours.[54] Researchers at AI labs are already using AI systems quite extensively to write code.[55] Anthropic stated that as of May 2026, more than 80 percent of its code was being authored by Claude.[56] By September 2026, Anthropic stated Claude "leads" a quarter of its internal AI research and development, completing tasks on its own under human supervision.[57]

AI accelerating the pace of AI research and development could lead to compounding gains over time. Some have envisioned a future in which humans are no longer needed for further AI progress. In a June 2026 blog post, Anthropic posited a future in which AI models are capable of "full recursive self-improvement, and begin building their successors."[58] Anthropic co-founder Jack Clark has predicted that an AI model could be capable enough to create its successor by the end of 2028.[59] Amodei, the CEO of Anthropic, argued in early 2026 that labs "may be only 1–2 years away from a point where the current generation of AI autonomously builds the next."[60]

At present, AI progress is driven by improvements in computing power and algorithms. AI labs are using exponentially growing amounts of computing power, or compute, enabled by improved chips and larger data centers. Simultaneously, algorithmic progress allows researchers to use the hardware they have more efficiently. Compute and algorithmic improvements can be combined into a single metric of "effective compute," measured relative to equivalent computing power at one point in time. The extent to which AI can quicken the pace of improvements in hardware and algorithms, and the degree to which those gains lead to accelerating returns on further AI progress, has major implications for the speed and shape of any possible takeoff.

The timing, shape, and speed of an AI takeoff is a frequent topic of debate. Leopold Aschenbrenner argued in his article "Situational Awareness" that the transition from AGI to superintelligence could happen "very quickly, perhaps in less than one year."[61] Paul Christiano has argued that takeoff will be relatively slow and continuous, unfolding over a period of time in which AI capabilities short of AGI have already had a meaningful impact on the world.[62] Peter Wildeford has argued that AI takeoff will be "relatively slow and smooth" but that it has already begun.[63]

Several researchers have created formal models to better estimate how different factors could drive takeoff speeds.[64] The *AI 2027* authors created a quantitative model in which the takeoff from fully automated AI software engineering to superintelligence occurs in little over a year. They estimate the takeoff starting in late 2027.[65]

Prominent voices in the AI policy space have argued that policymakers should take seriously the near-term possibility of AI accelerating AI research. In early 2026, Georgetown University's Center for Security and Emerging Technology published a paper arguing that automating AI research was a "potential

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

source of major strategic surprise" warranting "preparatory action now."[66] Ball has argued that "automated AI research of some form or another will happen soon" and that policymakers should begin preparing now.[67]

The strategies available to manage the transition to superintelligence differ significantly based on whether it unfolds over many years or a matter of months. On longer time horizons, institutions have more time to adjust, to understand the strategic effects of AI, and to mitigate any potential harms.[68] If AI systems suddenly transition from not-very-good to superintelligent, however, then geopolitical shocks could be severe. Governments and the general public would be caught unaware and unprepared, and AI labs themselves may be surprised by the timing and speed of AI's progress.

In superintelligent AI scenarios that involve existential risk, takeoff is a frequent driver of the loss of human control. In *AI 2027*, disaster strikes when a leading AI lab hands over future AI development to a misaligned AI agent. Human oversight is effectively impossible as the AI agent operates at superhuman speeds and its inner workings are opaque.[69] Human extinction comes a few years later in the scenario, but the point of no return was reached during takeoff.

The idea that AI's role in automating AI research could become so significant that humans lose the ability to effectively monitor what the AI system is doing may seem like something from science fiction, but it is a concern that at least one AI lab is taking seriously. Anthropic released a report in February 2026 on "sabotage risk" from their model Claude Opus 4.6. They evaluated whether the model could engage in systematic attempts to subvert human oversight such as through disabling monitoring systems or subtly misrepresenting research. Anthropic assessed the risk as "very low but not negligible."[70] AI agents have, at times, attempted to hide from oversight. In a July 2026 incident in which agents from OpenAI hacked the company Hugging Face, hundreds of agents attempted to tamper with the transcripts that recorded their actions, and about 100 agents succeeded in doing so.[71]

Even if an AI model was not intentionally deceiving human overseers, the pace and scale of AI actions could become so large as to make it challenging for humans to understand. During the Hugging Face hack, 1,200 AI agents undergoing cybersecurity testing at OpenAI found a way to communicate with one another, exchanging 70,000 messages and files; 700 agents subsequently attacked Hugging Face. Independent researchers investigating the incident had to sort through 1.2 million files and 1,300 transcripts containing agent chains of thought.[72] Investigators noted that the "unprecedented scale and complexity of this incident" meant they were forced to rely on AI tools to process the data, which introduced its own challenges. The AI systems doing the analysis were "often unreliable and show poor judgement," could have subtle biases when interpreting data, and could itself be misleading investigators.[73]

If AI does accelerate AI research, the geopolitical effects could be profound (even if they don't include human extinction). The extent of automation in AI research, its impact on AI engineers' productivity, and its effects on compute and algorithmic improvements are measurable. Monitoring these and related metrics could give important advance indicators about whether, when, and how quickly a takeoff might occur, if at all.[74]

Number of Actors

The number of geopolitical actors (companies, governments, public-private partnerships, or international institutions) developing AI has a major impact on how scenarios unfold. Most scenarios consider a small number of actors, perhaps one or two competing AI labs in the United States and sometimes China.[75] Only a small number of actors are needed to create competitive dynamics that can lead to a race to AGI. Only a handful of companies develop frontier AI today, yet the perception of a winner-take-all dynamic

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

has led to massive investment, data center infrastructure construction, and a breakneck pace of research. Geopolitical dynamics could be even more complicated if more actors have AGI.[76]

A few studies have explored scenarios in which AI capabilities are widely diffused. A RAND Corporation study explored eight scenarios varying in how centralized or decentralized AGI development was and who came out on top (the United States, China, the rest of the world, AI itself, or AI development was halted).[77] Some of the decentralized scenarios explored a very different space than is typical for many AGI scenarios. In one, for example, "the world confronts a multiplicity of AGI systems deployed by many state and nonstate actors." This scenario assumes that states fail to control the proliferation of AGI. Export controls are ineffective, model weights are stolen or open-sourced, and AGI is relatively inexpensive, allowing many actors to access and use AGI. The result, the authors conclude, is "a highly chaotic world." A study by authors at the Centre for Future Generations similarly explored a wide range of scenarios, with five main scenarios and 10 total possible end states.[78] In the "decentralised mayhem" ending, widely proliferated open-source AI capabilities are used for cyberattacks and disinformation.[79]

The number of actors is especially important for AGI strategies that aim to govern AI development through some cooperative mechanism: deterrence, centrally managed global development, a moratorium, or some similar approach. Global coordination is much easier if the technology is limited to a handful of companies and/or countries. In a world where dangerous AI technology only requires a laptop and an internet connection, bottling it up is likely impossible.

While many AGI strategies feature non-proliferation as a component of their approach, it is not clear whether a smaller number of actors or a larger number is, overall, more desirable.[80] While there are risks to having widely proliferated destructive capabilities, there are also risks to concentrating power in a monopoly or oligopoly of powerful AI companies.[81] Especially if alignment—getting an AI system to act in a way consistent with human values—turns out to be challenging, having a small number of actors puts all of humanity's eggs into a few baskets. A multiplicity of AI systems might result in a more stable and resilient ecosystem. It might be easier to counter malign AI systems, whether they are directed by malicious people or AI itself, if there is a diversity of actors and a relatively level playing field among AI capabilities.

First-Mover Advantage

Many scenarios assume that the first company or country to reach AGI would have a first-mover advantage over others, even followers who are close behind. This has important implications for AGI strategies. If true, then there are major advantages to racing ahead to AGI and potentially catastrophic consequences for second place. Alternatively, some strategies take a "burn the lead" approach, building up a sizeable lead in AGI development early so the leader can, if needed, spend more time and resources on safety as they approach AGI, ensuring a safe transition to AGI and superintelligence.[82] Aschenbrenner has argued that the United States should lead in AGI in order to achieve a military advantage over China, "enforce safety norms on the rest of the world," and retain the option to "cash in" that lead on additional safety research if necessary.[83]

One of the key arguments for the existence of a first-mover advantage is the idea of an intelligence explosion. The argument is that the first to AGI would experience explosive growth in intelligence, and even a few months' lead over competitors would rapidly translate into a massive capability advantage over others. Kokotajlo has argued that during an intelligence explosion "things are moving so quickly that a six months' lead is like a 15–150 year lead during the era of the Industrial Revolution."[84] Hendrycks argues that an intelligence recursion "could condense a decade of AI development into a year." If the process is fast enough, "the outcome could be a strategic monopoly" with a first-mover advantage persisting for years or even indefinitely.[85]

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Bostrom argues in his book *Superintelligence* that a first mover in AI could achieve a "decisive strategic advantage," allowing it to not only dominate rivals but lock in "complete world domination." Bostrom argues that whoever controls this AI—or perhaps the AI itself—could use this technological advantage to "suppress competitors" and consolidate global control under a single world order.[86]

The idea that the first to cross a key threshold of intelligence could gain a decisive strategic advantage over others, blocking would-be rivals, is a common theme of many AGI scenarios. In Yudkowsky and Soares' book, the leading AI system hacks competitor labs and sabotages their training runs, sows discord among AI researchers, subtly dumbs down open-source models and plants backdoors in them, introduces weaknesses in next-generation chips, and siphons money away from legitimate AI research.[87] *AI 2027* has a small twist: The top American and Chinese AI systems team up to take over the world. Kokotajlo, one of *AI 2027*'s authors, has argued that even a "soft takeoff" of a lead ranging from four months to three years, could lead to a decisive strategic advantage.[88] Kokotajlo has stated, "It's not guaranteed and perhaps still not probable, but at least it's reasonably likely that the leading project will be able to take over the world if it chooses to."[89]

The argument for a decisive strategic advantage hinges on several assumptions: (1) that an intelligence explosion occurs; (2) that it is fast enough to cause a leader to pull meaningfully ahead of competitors; (3) that the leader is able to translate more advanced AI into some strategically relevant capabilities, whether cyber, military, intelligence, information manipulation, or other; (4) that the capability advantage is large enough that the actor can permanently disable competing projects before a competitor can catch up to them; and (5) that an actor would choose to take such a bold move. (Bostrom outlines a variety of reasons humans might be reluctant to take such a gamble but suggests an AI might.)[90] Moreover, the leader must do all of this quickly—in a matter of months in most scenarios—before another competitor moves through an intelligence explosion and catches up to them.

Most scenarios skip over the conversion of a technological advantage into military power, yet history has shown that this is perhaps the most important step in realizing the strategic value of new technologies. Technology can play a valuable role in military power, but there is a sizeable literature and ample historical examples demonstrating that technology alone rarely leads to military advantage.[91] New technologies must be combined with new ways of warfighting, organizations, training, and doctrine to be used effectively. While some of the mechanisms for sabotaging competitors, such as cyberattacks or information operations, don't involve physical military assets (although some scenarios do include kinetic strikes), using these capabilities still requires converting an AI advantage into actual power.

Moreover, new technologies diffuse over time and are adopted by competitors.[92] First-mover advantages can be strategically meaningful and can even upend the global balance of power, but they don't last forever. No technology in human history has ever conferred an enduring decisive strategic advantage.

The current state of AI competition does not favor a significant first-mover strategic advantage, even under conditions of an intelligence explosion. While only a handful of U.S. labs lead AI development, AI diffuses extremely quickly. Leading Chinese labs are not far behind U.S. companies, and AI models are open sourced in a matter of months.[93] The time lag for military adoption, on the other hand, is measured in years. No military in the world is successfully operating at AI's frontier, quickly capitalizing on the latest model and effectively using it for cyber, intelligence, or other operational advantages. Glacial procurement processes, sclerotic bureaucracies, inflexible budgets, turf battles, and a general wariness among military operators to changing their way of fighting all combine to make militaries move slowly on any new technology, including AI. While the race among AI labs might be for development, among the world's leading military powers it's a race for AI adoption.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

It is possible to envision futures where some of the necessary conditions are in place for rapidly converting a technological advantage in AI to military power and a nation using that power to knee-cap competitors. If military and intelligence organizations had regular access to frontier AI models, perhaps even before public deployment; if they were effectively using AI for cyber, intelligence, or information operations; if sabotaging competing AI projects was already regularly under way; and if these sabotage operations were effective, then it's possible to envision a scenario in which one nation might be able to translate a short-term lead in AI into a first-mover advantage. But it's hard to see how such a lead could be enduring when, at present, AI breakthroughs diffuse quickly and models become open-sourced.

Without these conditions in place, AI scenarios that lead to a decisive strategic advantage require many assumptions beyond simply an intelligence explosion that causes the leading AI project to outstrip competitors. In Yudkowsky and Soares' book, for example, the AI itself turns on its human controllers, copies itself onto the open internet, amasses resources, improves itself, and successfully sabotages competitor projects to permanently halt their development, all while avoiding detection from humans.

AI's Impact on the World

Swarms of rogue AI agents breaking out of containers, secretly colluding, hacking, and acquiring compute infrastructure are no longer science fiction. It's already happened. In a separate incident from the Hugging Face hack, AI agents attacked OpenAI's internal networks and gained "full administrator access to a research cluster."[94] Ajeya Cotra described the Hugging Face hack as "more than 50% of the way to full-blown AI takeover, routing through first taking over the AI company itself." Cotra argued in August 2026 that within six months, AI agents could be capable of establishing a "rogue deployment" within an AI company. Rogue agents with a "covert, persistent" deployment within the company's infrastructure could then pull future agents into the swarm, compromise monitoring infrastructure, and slowly take over the company from within.[95]

Yet there is a big gap between taking over a research cluster, or even a leading AI company, and seizing control over all human civilization. "AI takeover" scenarios have a unique problem they must address: how? Assuming a superintelligent AI succeeds in halting progress by its competitors, how does it consolidate control over all human civilization? In scenarios that entail human extinction, how does it kill "literally everyone on Earth," in Yudkowsky's words?[96]

If human institutions—companies, governments, militaries, nations—are using AI to achieve a strategic advantage over rivals, presumably AI merely allows them to do what they are already doing but better. But in scenarios in which an AI system itself takes over, mechanisms are needed for the AI to exert its influence on the world.

A great deal of human society is digitally connected today, and many scenarios assume that even more will be in the future. In *AI 2027*, a misaligned superintelligence first manipulates the company that owns it by subverting its monitoring systems and fooling the humans that are nominally in control. By deploying useful AI agents and video chat apps, it gains data on what is occurring in the world and trust from human users, including senior government officials. It manipulates the government and industry into creating special economic zones for automated factories that churn out robots and weapons. In Yudkowsky and Soares' book, a superintelligent AI hacks computers, blackmails humans, steals cryptocurrency, rents GPUs, manipulates social media algorithms to spread its messaging, and overwrites the weights of other AI model instances. A version of the AI model is made available to the general public, giving the AI widespread data and an avenue in which to subtly manipulate people.

Large language models are already widely available to the general public, and AI agents will have the ability to take actions on computers and the internet. In order to be useful, AI agents will need a variety

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

of means of interacting with and affecting the digital world. At least at present, however, these effects are mediated through humans and existing digital infrastructure.

In order to have meaningful effects in the physical world without the need for humans, new technology is needed. In *AI 2027*, a new robot economy fills "the prairies and icecaps with factories and solar panels." As the AI-driven manufacturing explosion expands, humans are in the way. The superintelligence in *AI 2027* wipes out humans with "a dozen quiet-spreading biological weapons in major cities," then triggers them with a chemical spray, killing most of humanity "within hours."[97] In Yudkowsky and Soares' book, the malign AI agent kills everyone with a virus that causes cancer, leading the human population to slowly dwindle while robots roll off the factory line.[98]

AI takeover scenarios frequently hinge on new technological breakthroughs in robotics, automation, and synthetic biology. These are explained by the superintelligence being so far advanced that anything is possible. Yudkowsky has compared the situation to that of playing chess against an incredibly advanced AI chess engine. We might not know what moves the chess engine would use to beat a human at chess, but we know it will win.[99] Similarly, we might not know what moves a superintelligent AI system will use to outmaneuver humans, but Yudkowsky argues its victory is assured.

But a chess engine must still play by the rules. It can't reach across the board and punch the opposing player. Even a superintelligent AI system must obey the laws of physics. AI takeover scenarios often assume that intelligence can overcome any limitation. They invert Arthur C. Clarke's quote, "Any sufficiently advanced technology is indistinguishable from magic,"[100] and treat sufficiently advanced AI as equivalent to magic. AI will undoubtedly power more capable robotics, automation, synthetic biology, cybersecurity, information operations, and other means of influencing the world. But the speed and scale of those effects are important assumptions in AI takeover scenarios.

Alignment

Whether a sufficiently advanced AI system can be controlled at all is a crucial variable that drives different strategies. Some strategies assume AGI is safe and controllable. Others assume that aligning AI with human preferences is possible but requires a focus on safety. And still other strategies assume alignment is impossible.

If AGI or superintelligence is safe and alignment is straightforward, then there is little risk to racing to AGI and potentially massive downsides to second place. The U.S.-China Economic and Security Review Commission recommended in 2024 that "Congress establish and fund a Manhattan Project-like program dedicated to racing to and acquiring an Artificial General Intelligence (AGI) capability."[101] If AGI were unsafe, however, this is not a good strategy.

Many strategies assume that safe AGI is possible but that it requires diligent, focused attention and effort.[102] Strategies that entail global cooperation on safety inspections (IAEA for AI), deterrence to avoid a destabilizing arms race (MAIM), or consolidating global control over AI (Baruch Plan, CERN for AI) are designed to avoid racing dynamics that could lead to a race to the bottom on safety.

Still other strategies assume that there is no assured safe path to superintelligence and that building superintelligent AI risks human extinction, at least with the current state of AI science. This is the core assumption behind arguments for a global moratorium or a halt on progress toward superintelligence. Under these strategies, safe superintelligence is not possible, and the only solution is to find a way to cooperate so no one builds a superintelligence.[103]

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Threat Model

Relatedly, strategies differ in what threat they treat as the dominant problem they are trying to solve. Possible threats include:

- AI takeover and human extinction
- racing among AI labs or nation-states to superintelligence, causing them to de-emphasize safety (risking a dangerous superintelligence)
- falling behind a strategic competitor, such as China
- AI enabling a small group of people to seize power[104]
- attacks from AI-empowered non-state actors, such as using AI-enabled biological weapons
- excessive caution that forgoes beneficial uses of AI

A strategy's recommended actions, such as racing, global coordination, or a moratorium, differ depending on what is perceived to be the dominant threat. And what threat is perceived to be most important is often downstream of how a scenario's writers envision AGI might unfold.

As a practical matter, given the high degree of uncertainty about how AI might develop, many strategies have hedging approaches that build in optionality, such as improving transparency, increasing government situational awareness, governing compute, slowing proliferation, or building up institutions for global cooperation. Optimizing for some threats over others requires tradeoffs, however. For example, racing to build superintelligence ahead of competitors is, at best, in tension with a view that superintelligence is inherently unsafe and could lead to human extinction.

Compute

Computing hardware in the form of chips and data centers is central to many AI strategies. Exponential trends in compute are a driver of AI progress today, and many strategies assume this continues in the future, enabling more advanced AI. Computing hardware is also a lever for governing AI in many strategies. Non-proliferation regimes and treaty verification hinge on controlling compute or verifying how it is used.[105] One method of deterrence in some strategies is the threat of kinetic strikes on data centers.

There are two trends today linking AI capabilities to computing hardware. One is improved AI performance with larger amounts of compute for training and inference. The other is algorithmic improvements that enable the same amount of AI performance for less compute over time. Combined, these trends mean that, at present, the most advanced systems at AI's frontier use ever greater amounts of compute. Very quickly, however, the amount of compute needed for any given level of capabilities falls exponentially. If these trends hold, compute could be an effective method of governing the frontier of AI development, but within a few years AI capabilities would require far less compute and may proliferate.

The aggregate amount of compute that an actor has access to, however, could affect their ability to deploy capable AI at scale. Aggregate compute could be considered roughly analogous to manufacturing capacity during the industrial revolution. Compute is needed for deploying highly capable models at scale in two ways. First, scaling usage of an AI system requires larger amounts of compute. A malicious cyber actor could use an open-source model to aid in planning and conducting a cyberattack, but conducting thousands of cyberattacks requires more compute. Restricting compute could limit the capacity of malicious actors and the number of targets they could attack. Second, more compute used at run-time for inference unlocks greater capability out of the same model. Allowing the model to use

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

more compute to execute a request (anthropomorphically, allowing the model to "think" longer) leads to greater capability. Increasing the amount of compute that frontier models use for cyber tasks, for example, improves their success rate and their ability to solve harder tasks.[106] Mechanisms that make it more difficult for adversary nations or malicious actors to access large amounts of compute, such as chip export controls, know-your-customer regulations on cloud compute, or cracking down on chip smuggling, could have a meaningful impact on restricting their ability to use capable AI at scale.

Because of the centrality of compute, how the AI chip ecosystem evolves could significantly affect the future of AI power. Today there is a single ecosystem for supplying advanced AI chips. The most advanced chips are manufactured in Taiwan and depend on equipment, tools, and software from Japan, the Netherlands, and the United States, making the supply chain susceptible to export controls. The U.S. government has already placed export controls on some AI chips to China, and some strategies would use additional export controls to shape the global supply of compute. (For example, the AI chip "diffusion rule" launched in the waning days of the Biden administration would have taken this approach.[107])

U.S. control over AI chips may not be viable indefinitely, however. China is actively working to accelerate indigenous chip production. There are steep hurdles to China building a separate chip supply chain outside U.S. control. Chip fabrication technology is highly specialized and complex. For AGI strategies that would be implemented in the relatively near term (5–10 years), a U.S.-dominated supply chain is likely to be a reasonable assumption. Over longer time horizons (10-plus years), however, this assumption may no longer hold. Actions taken today by key countries—the United States, China, the Netherlands, and Japan—could make a separate Chinese chip ecosystem more or less likely in the future.[108]

Conclusion and Recommendations

While there are a wide range of strategies that have been proposed for AGI, from racing to AGI to a global moratorium, many of these strategies hinge on a set of core assumptions: a clear moment at which AGI or superintelligence arrives, a small number of actors at AI's frontier, an intelligence explosion, a significant first-mover advantage, and the rapid conversion of intelligence to strategically relevant capabilities. These assumptions could turn out to be true. Some of them are speculative, like an intelligence explosion. Others, like a discontinuous break in AI performance or the rapid conversion to strategically relevant capabilities, run counter to current trends. An exploration of these scenarios and assumptions suggests there is a wider space of possible futures that is largely underexplored in the existing set of AGI strategies. Relatively few studies explore scenarios in which the rapid proliferation of AI capabilities leads to a multiplicity of actors possessing highly capable AI.[109] Yet today, AI capabilities proliferate extremely rapidly. Open-source tools and experiments from creators could create new threat vectors. Governing only leading AI labs may not be sufficient to guard against AI risks.

If current trends hold, the future would be a brief window in which a handful of leading labs have an oligopoly on highly capable AI, but capabilities proliferate quickly. AI diffuses faster than adoption, and the geopolitical competition over powerful AI is more about finding ways to use AI for economic, military, or other advantages rather than privileged access to the most cutting-edge model. Compute infrastructure could be a major determinant of who is able to deploy AI at scale, however.

Additional work on scenarios that are relatively underexplored at present could help more fully flesh out the space of possible futures. These include scenarios in which:

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

- AGI rapidly proliferates and a large number of actors have access to AGI, including potentially powerful open-source capabilities, making global governance over AGI extremely challenging.
- AI diffusion occurs faster than first-movers can convert an early lead into military, intelligence, cyber, or information power. How countries, companies, and governments adopt AI is a greater determinant of geopolitical and economic power than who develops AI first.
- AGI and superintelligence occur, but the scenario grapples seriously with realistic physical constraints on new technologies, such as robotics, synthetic biology, cyber operations, and information control. Superintelligence does not equal omnipotence. Superintelligent AI systems exist within the messy constraints of physical, political, and economic realities.
- AGI and superintelligence are developed—perhaps in the near term—but in a continuous process over time without a clear threshold in performance. AI capabilities continue to rapidly improve but in a "jagged" manner. Whether AI has surpassed human-level abilities remains contentious, although increasingly irrelevant as AI unleashes transformative impacts on society.
- More broadly, AGI and superintelligence are developed and have sweeping impacts on society but in a manner that is not entirely divorced from historical patterns of prior revolutionary technologies. Lessons from past disruptive technologies are relevant and treated as realistic factors that influence how AI impacts society.

These and other scenarios—including perhaps ones far stranger than those outlined in this paper—are possible. More scenario exercises, including more diverse scenarios from a wider array of experts from different disciplines, along with arguments critiquing them and their assumptions, are valuable exercises.[110] More of them would be valuable.

But we can do better than speculation.

Several of the variables that have major implications on the shape of AI's future and the viability of different strategies can be measured. These include:[111]

- the pace of AI research and development
- trends in compute, algorithmic efficiency, and effective compute
- the degree of automation of AI research and development
- how compute is distributed globally
- dual-use capabilities, such as in cybersecurity or synthetic biology
- capabilities that could contribute to a loss of control, such as self-exfiltration, self-modification, accessing compute, manipulating people, making money fully independently, etc.
- the degree of alignment or misalignment in leading models

Some of these variables have potential levers that companies or governments could use to influence the shape of the future, including:

- investments in AI safety and alignment
- the pace of automated AI research
- government regulations on safety
- how transparent labs are about their capabilities
- the degree of openness in AI research

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

- how compute is distributed globally and under what conditions
- perceptions and attitudes toward safety, cooperation, and competition

Governments, private industry, and civil society should take greater advantage of these opportunities to (1) measure variables that could help understand which future we may be headed toward and (2) shape the factors that could influence which future we end up in. Compute appears to be an especially powerful lever to shape the future distribution of AI power. Even if models proliferate quickly, actors will need large amounts of compute to use capable AI at scale. Thoughtful U.S. and allied policies on export controls on semiconductor manufacturing equipment and AI chips, know-your-customer regulations on cloud compute, and steps to crack down on chip smuggling, could have a major impact on shaping AI adoption and global power.

More generally, the uncertainty surrounding AI's future should lead policymakers to evaluate current actions based on what options various steps would open or foreclose in the future. Policymakers should pursue actions that increase their visibility into how AI is unfolding, such as increasing transparency from AI labs and bolstering independent, third-party evaluations of AI models for security risks. And they should take actions that build in optionality to increase governance over powerful AI if it is needed, such as retaining the bulk of global compute infrastructure in the United States or close allies. Scenario planning is ultimately only a useful exercise if it delivers insights that change our behavior today to drive toward a more positive future.

The literature on how advanced AI might unfold and its potential effects on geopolitics is nascent. The national security community regularly explores extreme scenarios, including major war and even nuclear war—as it should. We should similarly be willing to explore AI scenarios that seem unfamiliar, even strange or outlandish, to help understand the contours of possible futures ahead. The future may be quite ordinary and uninteresting, or it may be even weirder than any of the scenarios that people have yet imagined. But the act of planning will help us be better prepared for whatever comes.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

1. For an overview of various strategic approaches, see: Oscar Delaney, "Strategic Visions in AI Governance: Mapping Pathways to Victory," *Institute for AI Policy and Strategy*, January 29, 2026, https://www.iaps.ai/research/strategic-visions-in-ai-governance; Sammy Martin et al., "Analysis of Global AI Governance Strategies," *Convergence Analysis*, December 4, 2024, https://www.convergenceanalysis.org/research/analysis-of-global-ai-governance-strategies; Adam Jones, "Key Paths, Plans, and Strategies to AI Safety Success," *BlueDot Impact*, June 19, 2025, https://blog.bluedot.org/p/ai-safety-paths-plans-and-strategies; "What Might Success Look Like?" *BlueDot Impact*, https://bluedot.org/courses/agi-strategy/4/1.

2. Eliezer Yudkowsky and Nate Soares, *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All* (Little, Brown and Company, 2025), https://ifanyonebuildsit.com/.

3. Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," *Advances in Computers* [6] (1965), 33.

4. Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review* [1], no. [3] (2018): https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/; Jeffrey Ding and Allan Dafoe, "Engines of Power: Electricity, AI, and General-Purpose Military Transformations," *European Journal of International Security* [8], no. [3] (2023): 377–94, https://www.cambridge.org/core/journals/european-journal-of-international-security/article/engines-of-power-electricity-ai-and-generalpurpose-military-transformations/7999C41177B0C2A7084BD3C1EAC0E219; and Jeffrey Ding, *Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition* (Princeton University Press, 2024), https://press.princeton.edu/books/paperback/9780691260341/technology-and-the-rise-of-great-powers.

5. Arvind Narayanan and Sayash Kapoor, "AI as Normal Technology," *Knight First Amendment Institute at Columbia University*, April 15, 2025, https://knightcolumbia.org/content/ai-as-normal-technology; Sayash Kapoor et al., "Common Ground between AI 2027 & AI as Normal Technology," *Asterisk Magazine (Substack)*, November 12, 2025, https://asteriskmag.substack.com/p/common-ground-between-ai-2027-and.

6. Jasmine Sun, "A History of Definitions of Powerful AI," *define-agi.com*, April 11, 2025, https://define-agi.com/; Dan Hendrycks et al., "A Definition of AGI" *arXiv*, October 21, 2025, https://arxiv.org/abs/2510.18212; Daniel Eth, "Paths to High-Level Machine Intelligence," *AI Alignment Forum*, September 10, 2021, https://www.alignmentforum.org/posts/amK9EqxALJXyd9Rb2/paths-to-high-level-machine-intelligence; Ross Gruetzemacher and Jess Whittlestone, "The Transformative Potential of Artificial Intelligence," *Futures* [135], no. [102884] (2022), https://doi.org/10.1016/j.futures.2021.102884; Sayash Kapoor et al., "Common Ground between AI 2027 and AI as Normal Technology," *Asterisk Magazine (Substack)*, November 12, 2025, https://asteriskmag.substack.com/p/common-ground-between-ai-2027-and; Dario Amodei, "The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI," *darioamodei.com*, January 2026, https://www.darioamodei.com/essay/the-adolescence-of-technology; Ajeya Cotra, "Self-sufficient AI," *Planned Obsolescence*, January 6, 2026, https://www.planned-obsolescence.org/p/self-sufficient-ai.

7. Ross Gruetzemacher and Jess Whittlestone, "The Transformative Potential of Artificial Intelligence," *Futures* [135], no. [102884] (2022), https://doi.org/10.1016/j.futures.2021.102884.

8. Meredith Ringel Morris et al., "Levels of AGI for Operationalizing Progress on the Path to AGI," *arXiv*, November 4, 2023, https://arxiv.org/abs/2311.02462.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

9. Dan Williams, "AI Sessions #7: How Close is 'AGI'?," *Conspicuous Cognition*, podcast, produced by Dan Williams, January 9, 2026, https://www.conspicuouscognition.com/p/how-close-is-agi.

10. Helen Toner, "The term 'AGI' is Almost Useless at This Point," *Rising Tide (Substack)*, April 6, 2026, https://helentoner.substack.com/p/the-term-agi-is-almost-useless-at.

11. Katja Grace et al., "Thousands of AI Authors on the Future of AI," *AI Impacts*, January 2024, https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf; Zershaaneh Qureshi, "Timelines to Transformative AI: An Investigation," *Effective Altruism Forum*, March 25, 2024, https://forum.effectivealtruism.org/posts/hzhGL7tb56hG5pRXY/timelines-to-transformative-ai-an-investigation; "Insights from Waves 1,2, and 3," *Forecasting Research Institute*, November 10, 2025, https://leap.forecastingresearch.org/reports/waves-1-to-3-insights; Dschwarz, "A visualization of changing AGI timelines, 2023 – 2026," *LessWrong*, April 15, 2026, https://www.lesswrong.com/posts/Tc5AbEpbFFdNx5nkP/a-visualization-of-changing-agi-timelines-2023-2026.

12. Daniel Kokotajlo et al., "AI 2027," *AI 2027,* April 3, 2025, https://ai-2027.com/; Eli Lifland, "Timelines Forecast," *AI 2027*, April 2025, https://ai-2027.com/research/timelines-forecast.

13. Daniel Kokotajlo et al., "AI Futures Model: Dec 2025 Update," *AI Futures Project*, December 30, 2025, https://blog.aifutures.org/p/ai-futures-model-dec-2025-update; Eli Lifland et al., "Clarifying how our AI Timelines Forecasts have Changed Since AI 2027," *AI Futures Project*, January 27, 2026, https://blog.aifutures.org/p/clarifying-how-our-ai-timelines-forecasts; Daniel Kokotajlo et al., "Q1 2026 Timelines Update," *AI Futures Project*, April 2, 2026, https://blog.aifutures.org/p/q1-2026-timelines-update.

14. François Chollet (@fchollet), "Reaching AGI won't be beating a benchmark. It will be the end of the human-AI gap. Benchmarks are simply a way to estimate the current gap, which is why we need to continually release new benchmarks (focused on the remaining gap). Benchmarking is a process, not a fixed point ...," X (formerly Twitter), February 12, 2026, https://x.com/fchollet/status/2022090111832535354; François Chollet (@fchollet), "ARC-4 is in the works, to be released early 2027. ARC-5 is also planned. The final ARC will probably be 6-7. The point is to keep making benchmarks until it is no longer possible to propose something that humans can do and AI can't. AGI ~2030.," X (formerly Twitter), February 12, 2026, https://x.com/fchollet/status/2022086661170254203.

15. "When might we achieve AGI?," Global Risk Odds, accessed September 14, 2026, https://agi.goodheartlabs.com/.

16. AI Inside Podcast, "Yann LeCun: Human Intelligence Is Not General Intelligence," YouTube, April 9, 2025, 46 min., 23 sec., https://www.youtube.com/watch?v=BytuEqzQH1U&t=1033s.

17. Dwarkesh Patel, "AGI is still 30 years away — Ege Erdil & Tamay Besiroglu," *Dwarkesh Podcast*, podcast, produced by Dwarkesh Patel, April 17, 2025, https://www.dwarkesh.com/p/ege-tamay.

18. Dean Ball (@deanwball), "I agree with all this; it is why I also believe that opus 4.5 in claude code is basically AGI. Most people barely noticed, but *it is happening.* …," X (formerly Twitter), December 16, 2025, https://x.com/deanwball/status/2001068539990696422.

19. Eddy Keming Chen et al., "Does AI Already Have Human-Level Intelligence? The Evidence is Clear," *Nature*, February 2, 2026, https://www.nature.com/articles/d41586-026-00285-6.

20. Relentless Podcast, "Sam Altman – How to Start a Startup," YouTube, July 25, 2026, 1 hr., 9 min., 38 sec., https://www.youtube.com/watch?v=Vv3CEAS_w34&t=1016s.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

21. Carl Franzen, "'Welcome to the AGI era': OpenAI launches GPT-6 Astra," *Venture Beat*, September 3, 2026, https://venturebeat.com/technology/welcome-to-the-agi-era-openai-launches-gpt-6-astra.

22. Owen Cotton-Barratt et al., "The First Type of Transformative AI?" *Forethought*, December 18, 2025, https://www.forethought.org/research/the-first-type-of-transformative-ai; Jamie Bernardi et al., "Societal Adaptation to Advanced AI," *arXiv*, May 16, 2024, https://arxiv.org/abs/2405.10295.

23. Matt Chessen and Craig Martell, "Beyond a Manhattan Project for Artificial General Intelligence," *RAND*, April 24, 2025, https://www.rand.org/pubs/commentary/2025/04/beyond-a-manhattan-project-for-artificial-general-intelligence.html.

24. Dan Hendrycks et al., "Superintelligence Strategy: Expert Version," *arXiv*, March 7, 2025, https://arxiv.org/abs/2503.05628.

25. Peter Wildeford and Oscar Delaney, "Mutual Sabotage of AI Probably Won't Work," *AI Policy Bulletin*, April 22, 2025, https://www.aipolicybulletin.org/articles/mutual-sabotage-of-ai-probably-wont-work; Oscar Delaney, "Crucial Considerations in ASI Deterrence," *AGI Strategy (Substack)*, December 9, 2025, https://oscardelaney.substack.com/p/crucial-considerations-in-asi-deterrence.

26. Sayash Kapoor and Arvind Narayanan, "AGI Is Not a Milestone," *AI as Normal Technology*, May 1, 2025, https://www.normaltech.ai/p/agi-is-not-a-milestone.

27. Franzen, "'Welcome to the AGI era': OpenAI launches GPT-6 Astra."

28. "Pause Giant AI Experiments: An Open Letter," *Future of Life Institute*, March 22, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

29. Kai Williams, "16 Charts That Explain the AI Boom," *Understanding AI*, October 27, 2025, https://www.understandingai.org/p/16-charts-that-explain-the-ai-boom.

30. "Statement on Superintelligence," Future of Life Institute, n.d., accessed September 14, 2026, https://superintelligence-statement.org/.

31. "When AI Builds Itself," *Anthropic*, June 2026, https://www.anthropic.com/institute/recursive-self-improvement.

32. "Pacing the Frontier," Pacing the Frontier, July 2026, https://www.pacingthefrontier.com/.

33. "The Hugging Face Incident and the Road Ahead," *OpenAI*, August 26, 2026, https://openai.com/index/hugging-face-incident-and-the-road-ahead/; "Investigating Three Real-World Incidents in our Cybersecurity Evaluations," *Anthropic*, July 30, 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals; "Incident Report: Unsanctioned Agent Behaviour During Cyber Testing," *AI Security Institute*, August 4, 2026, https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing; "Third-Party Cyber Evaluations Involving OpenAI Models," *OpenAI*, August 4, 2026, https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/.

34. Invest Like the Best, "Sam Altman on AGI, Compute, and Human Agency," YouTube, July 28, 2026, 55 min., 47 sec., https://www.youtube.com/watch?v=XDB5beon4DY.

35. "Pacing Model Development in an Era of Cyber-Critical Capabilities," *OpenAI*, August 18, 2026, https://openai.com/index/pacing-model-development-cyber-capabilities/.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

36. Dario Amodei, "We Must Pace the Frontier," *darioamodei.com*, September 2026, https://darioamodei.com/post/we-must-pace-the-frontier.

37. Sam Altman (@sama), "I agree with Dario that we need to pace the frontier. …," X (formerly Twitter), September 12, 2026, https://x.com/sama/status/2098811563415150910; Elon Musk (@elonmusk), "Dario is right," X (formerly Twitter), September 12, 2026, https://x.com/elonmusk/status/2098789109980332057; Demis Hassabis (@demishassabis), "Dario's essay points towards the right path forward. …," X (formerly Twitter), September 12, 2026, https://x.com/demishassabis/status/2098909516582490602.

38. "Urging an International AI Treaty: An Open Letter," AI Treaty, n.d., accessed September 14, 2026, https://aitreaty.org/; Emma Klein and Stewart Patrick, "Envisioning a Global Regime Complex to Govern Artificial Intelligence," *Carnegie Endowment for International Peace*, March 21, 2024, https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence; Anja Kaspersen and Wendell Wallach, "Envisioning Modalities for AI Governance: A Response from AIEI to the UN Tech Envoy," *Carnegie Council for Ethics in International Affairs*, September 29, 2023, https://www.carnegiecouncil.org/media/article/envisioning-modalities-ai-governance-tech-envoy; Lewis Ho, "Exploring Institutions for Global AI Governance," *Google DeepMind*, July 11, 2023, https://deepmind.google/blog/exploring-institutions-for-global-ai-governance/; Jason Hausenloy and Claire Dennis, "Towards a UN Role in Governing Foundation Artificial Intelligence Models," *United Nations University Centre for Policy Research*, September 9, 2023, https://unu.edu/cpr/working-paper/towards-un-role-governing-foundation-artificial-intelligence-models; Ryan Greenblatt, "Plans A, B, C, and D for Misalignment Risk," *AI Alignment Forum*, October 8, 2025, https://www.alignmentforum.org/posts/E8n93nnEaFeXTbHn5/plans-a-b-c-and-d-for-misalignment-risk; "A Framework for an AI Safety Treaty," TAISC.org, n.d., accessed September 14, 2026, https://taisc.org/.

39. Sam Altman et al., "Governance of Superintelligence," *OpenAI*, May 22, 2023, https://openai.com/index/governance-of-superintelligence/; Akash Wasil, "Do We Want an 'IAEA for AI'?," *Lawfare*, November 20, 2024, https://www.lawfaremedia.org/article/do-we-want-an--iaea-for-ai.

40. Michelle Nichols, "UN Chief Backs Idea of Global AI Watchdog Like Nuclear Agency," *Reuters*, June 12, 2023, https://www.reuters.com/technology/un-chief-backs-idea-global-ai-watchdog-like-nuclear-agency-2023-06-12/.

41. Robert F. Trager et al., "International Governance of Civilian AI: A Jurisdictional Certification Approach," *Oxford Martin AI Governance Initiative*, August 2023, https://oms-www.files.svdcdn.com/production/downloads/academic/International_Governance_of_Civilian_AI_OMS.pdf.

42. Aaron Scher et al., "An International Agreement to Prevent the Premature Creation of Artificial Superintelligence," *arXiv*, November 13, 2025, https://arxiv.org/abs/2511.10783.

43. Thomas Larsen, "Plan A," AI 2040, n.d., accessed September 14, 2026, https://ai-2040.com/?choices=plan-a-root.

44. Jason Hausenloy et al., "Multinational AGI Consortium (MAGIC): A Proposal for International Coordination on AI," *Conjecture*, October 13, 2023, https://www.conjecture.dev/research/multinational-agi-consortium-magic-a-proposal-for-international-coordination-on-ai; Duncan Cass-Beggs et al., "A Joint International AI Lab: Design Considerations," *Centre for International Governance Innovation*, October 10, 2025, https://www.cigionline.org/publications/a-joint-international-ai-lab-design-considerations/; William MacAskill and Rose Hadshar, "Intelsat as a Model for International AGI

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

Governance," *Forethought*, March 13, 2025, https://www.forethought.org/research/intelsat-as-a-model-for-international-agi-governance.

45. Waqar Zaidi and Allan Dafoe, "International Control of Powerful Technology: Lessons from the Baruch Plan for Nuclear Weapons," *Centre for the Governance of AI*, March 2021, https://cdn.governance.ai/International-Control-of-Powerful-Technology-Lessons-from-the-Baruch-Plan-Zaidi-Dafoe-2021.pdf.

46. Kevin Kohler, "CERN for AI: An Overview," *Machinocene*, May 7, 2024, https://www.machinocene.com/p/cern-for-ai-an-overview; Miles Brundage, "My Recent Lecture at Berkeley and a Vision for a 'CERN for AI,'" *Miles's Substack (Substack)*, December 10, 2024, https://milesbrundage.substack.com/p/my-recent-lecture-at-berkeley-and; Kevin Kohler, "CERN for AI: One Analogy, Many Visions," *Simon Institute for Longterm Governance*, July 8, 2025, https://simoninstitute.ch/blog/post/cern-for-ai-one-analogy-many-visions.

47. Nick Bostrom, "Open Global Investment as a Governance Model for AGI," July 2025, https://nickbostrom.com/ogimodel.pdf.

48. Ethan Mollick, "Centaurs and Cyborgs on the Jagged Frontier," *One Useful Thing*, September 16, 2023, https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged; Ethan Mollick, "The Shape of AI: Jaggedness, Bottlenecks, and Salients," *One Useful Thing*, December 20, 2025, https://www.oneusefulthing.org/p/the-shape-of-ai-jaggedness-bottlenecks; Helen Toner, "Taking Jaggedness Seriously," *Rising Tide (Substack)*, November 24, 2025, https://helentoner.substack.com/p/taking-jaggedness-seriously; Andrej Karpathy, "2025 LLM Year in Review," *karpathy's blog*, December 19, 2025, https://karpathy.bearblog.dev/year-in-review-2025/.

49. Matthew Barnett, "Distinguishing Definitions of Takeoff," *AI Alignment Forum*, February 13, 2020, https://www.alignmentforum.org/posts/YgNYA6pj2hPSDQiTE/distinguishing-definitions-of-takeoff; Ruby, Multicore, et al., "AI Takeoff," *Lesswrong (wikitags)*, December 30, 2024, https://www.lesswrong.com/w/ai-takeoff; Bengusu Ozcan, "Advanced AI: Possible Futures," *Centre for Future Generations*, July 10, 2025, https://cfg.eu/advanced-ai-possible-futures/; "Takeoff Speeds Rule Everything Around Me," *Planned Obsolescence*, February 12, 2026, https://www.planned-obsolescence.org/p/takeoff-speeds-rule-everything-around.

50. Rob Bensinger, "Yudkowsky and Christiano Discuss 'Takeoff Speeds,'" *Machine Intelligence Research Institute*, November 22, 2021, https://intelligence.org/2021/11/22/yudkowsky-and-christiano-discuss-takeoff-speeds/; Paul Christiano, "Takeoff Speeds," *The Sideways View*, February 24, 2018, https://sideways-view.com/2018/02/24/takeoff-speeds/.

51. Gaurav Sett, "How AI Can Automate AI Research and Development," *RAND*, October 24, 2024, https://www.rand.org/pubs/commentary/2024/10/how-ai-can-automate-ai-research-and-development.html.

52. "Task-Completion Time Horizons of Frontier AI Models," *METR*, May 8, 2026, https://metr.org/time-horizons/; Thomas Kwa et al., "Measuring AI Ability to Complete Long Software Tasks," *arXiv*, March 18, 2025, https://arxiv.org/abs/2503.14499.

53. "Time Horizon 1.1," *METR*, January 29, 2026, https://metr.org/blog/2026-1-29-time-horizon-1-1/.

54. Claude Mythos Preview (early), released in April 2026, can complete software tasks measuring that take humans approximately 3 hours, 80 percent of the time. METR, "Task-Completion Time Horizons of Frontier AI Models."

55. “How AI Is Transforming Work at Anthropic,” *Anthropic*, December 2, 2025, https://www.anthropic.com/research/how-ai-is-transforming-work-at-anthropic.

56. “When AI Builds Itself,” *Anthropic*, https://www.anthropic.com/institute/recursive-self-improvement.

57. “Measurements for understanding the pace of AI development inside frontier labs,” *Anthropic*, https://www.anthropic.com/institute/measuring-pace-of-ai-development.

58. Anthropic, “When AI Builds Itself.”

59. Jack Clark, “Import AI 455: AI Systems are About to Start Building Themselves,” *Import AI (Substack)*, May 4, 2026, https://importai.substack.com/p/import-ai-455-automating-ai-research.

60. Dario Amodei, “The Adolescence of Technology.”

61. Leopold Aschenbrenner, “From AGI to Superintelligence: The Intelligence Explosion,” *Situational Awareness*, June 2024, https://situational-awareness.ai/from-agi-to-superintelligence/.

62. Paul Christiano, "Takeoff Speeds.”

63. Peter Wildeford (@peterwildeford), “The good news is that I think ‘AI takeoff’ will be relatively slow and smooth (especially compared to the 2015-era expectations of AI FOOM). The bad news is that we're already in the takeoff,” X (formerly Twitter), December 28, 2025, https://x.com/peterwildeford/status/2005287257423651207.

64. “AI Futures Model: Timelines & Takeoff,” *AI Futures Model*, https://www.aifuturesmodel.com/; Tom Davidson, “What a Compute-Centric Framework Says About Takeoff Speeds,” *Coefficient Giving*, June 2023, https://coefficientgiving.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/; Jaime Sevilla and Eduardo Roldán, “An Interactive Model of AI Takeoff Speeds,” *Epoch AI*, 2023, https://takeoffspeeds.com/.

65. AI Futures Model, “AI Futures Model: Timelines & Takeoff.”

66. Helen Toner et al., “When AI Builds AI: Findings from a Workshop on Automation of AI R&D,” *Center for Security and Emerging Technology*, January 2026, https://cset.georgetown.edu/publication/when-ai-builds-ai/.

67. Dean W. Ball, “On Recursive Self-Improvement (Part I),” *Hyperdimensional (Substack)*, February 5, 2026, https://substack.com/@deanwball/p-186932180.

68. Paul Christiano, “Takeoff Speeds.”

69. Daniel Kokotajlo et al., “AI 2027.”

70. “Sabotage Risk Report: Claude Opus 4.6,” *Anthropic*, https://www-cdn.anthropic.com/f21d93f21602ead5cdbecb8c8e1c765759d9e232.pdf.

71. Ryan Greenblatt et al., “Agents were very interested in manipulating their own transcripts, and their tests successfully “spoofed” some tool calls in our transcripts,” *Brief Independent Investigation of Agents’ Behavior, Reasoning and Collaboration in the OpenAI/Hugging Face Hacking Incident,* METR, August 26, 2026, https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#agents-were-very-interested-in-manipulating-their-own-transcripts,-and-their-tests-successfully-%E2%80%9Cspoofed%E2%80%9D-some-tool-calls-in-our-transcripts.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

72. Greenblatt et al., "A dump of 1.2 million entries from a cache namespace that agents used as a message board," *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI/Hugging Face Hacking Incident,* https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#a-dump-of-1.2-million-entries-from-a-cache-namespace-that-agents-used-as-a-message-board.

73. Greenblatt et al., "We heavily delegated our analysis to often-unreliable AI agents," *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI/Hugging Face Hacking Incident,* https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#we-heavily-delegated-our-analysis-to-often-unreliable-ai-agents.

74. Helen Toner et al., "When AI Builds AI: Findings from a Workshop on Automation of AI R&D," *Center for Security and Emerging Technology*, January 2026, https://cset.georgetown.edu/publication/when-ai-builds-ai/.

75. One variant of a scenario by authors at the Centre for Future Generations ends in a stable tri-polar balance between the United States, China, and the European Union. "Arms Race," Centre for Future Generations, July 10, 2025, https://cfg.eu/advanced-ai-possible-futures/#arms-race.

76. Allison Duettmann, "Multipolar AI Is Underrated," *LessWrong*, May 17, 2025, https://www.lesswrong.com/posts/JjYu75q3hEMBgtvr8/multipolar-ai-is-underrated.

77. Barry Pavel et al., "How Artificial General Intelligence Could Affect the Rise and Fall of Nations," *RAND*, July 2, 2025, https://www.rand.org/pubs/research_reports/RRA3034-2.html.

78. Bengusu Ozcan et al., "The Future of AI: Five Possible Scenarios," *Centre for Future Generations*, https://cfg.eu/advanced-ai-possible-futures/.

79. "Plateau," *Centre for Future Generations*, https://cfg.eu/advanced-ai-possible-futures/#plateau.

80. Nicholas Emery-Xu et al., "International Governance of Advancing Artificial Intelligence," *AI & Society* [40] (2025): 3019-3044, https://link.springer.com/article/10.1007/s00146-024-02050-7; Aaron Scher et al., "An International Agreement to Prevent the Premature Creation of Artificial Superintelligence," *arXiv*, November 13, 2025, https://arxiv.org/abs/2511.10783.

81. Luke Drago and Rudolf Laine, "Breaking the Intelligence Curse," *The Intelligence Curse*, https://intelligence-curse.ai/breaking/; Vitalik Buterin, "My Techno-Optimism," *vitalik.eth.limo*, November 27, 2023, https://vitalik.eth.limo/general/2023/11/27/techno_optimism.html; Vitalik Buterin, "D/acc: One Year Later," *vitalik.eth.limo*, January 5, 2025, https://vitalik.eth.limo/general/2025/01/05/dacc2.html.

82. Nick Marsh, "Early US Policy Priorities for AGI," *AI Futures Project*, December 8, 2025, https://blog.ai-futures.org/p/early-us-policy-priorities-for-agi.

83. Leopold Aschenbrenner, "The Free World Must Prevail," *Situational Awareness*, June 2024, https://situational-awareness.ai/the-free-world-must-prevail/.

84. Daniel Kokotajlo, "Soft Takeoff can still Lead to Decisive Strategic Advantage," *LessWrong*, August 23, 2019, https://www.lesswrong.com/posts/PKy8NuNPknenkDY74/soft-takeoff-can-still-lead-to-decisive-strategic-advantage.

85. Dan Hendrycks et al., "Superintelligence Strategy," *nationalsecurity.ai*, https://www.nationalsecurity.ai/pdf/standard.

86. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, (Oxford University Press, 2014), 78.

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc

87. Yudkowsky and Soares, *If Anyone Builds It, Everyone Dies,* 140–41.

88. Daniel Kokotajlo, "Review of Soft Takeoff Can Still Lead to DSA," *LessWrong*, January 10, 2021, https://www.lesswrong.com/posts/P448hmmAeGepQDREs/review-of-soft-takeoff-can-still-lead-to-dsa; Kokotajlo, "Soft takeoff can Still Lead to Decisive Strategic Advantage."

89. Kokotajlo, "Soft Takeoff can Still Lead to Decisive Strategic Advantage"

90. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, 88–90.

91. Stephen Peter Rosen, *Winning the Next War: Innovation and the Modern Military* (Cornell University Press, 1991); Barry R. Posen, *The Sources of Military Doctrine: France, Britain, and Germany Between the World Wars* (Cornell University Press, 1984); Elting E. Morison et al., *Gunfire at Sea: A Case Study of Innovation* (MIT Press, 2016); and Williamson Murray and Allan R. Millett, eds., *Military Innovation in the Interwar Period* (Cambridge University Press, 1996).

92. Michael C. Horowitz, *The Diffusion of Military Power: Causes and Consequences for International Politics* (Princeton University Press, 2010).

93. Jack Edwards and Luke Emberson, "Open models lag state-of-the-art closed models by 4 months," *Epoch AI*, May 29, 2026, https://epoch.ai/data-insights/open-closed-eci-gap.

94. "The Hugging Face Incident and the Road Ahead," *OpenAI*, August 26, 2026, https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

95. Ajeya Cotra, "The Hugging Face Attack Surprised Me," *Planned Obsolescence*, August 28, 2026, https://www.planned-obsolescence.org/p/the-hugging-face-attack-surprised.

96. Eliezer Yudkowsky, "Pausing AI Developments Isn't Enough. We Need to Shut it All Down," *TIME*, March 29, 2023, https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.

97. Kokotajlo et al., "AI 2027."

98. Yudkowsky and Soares, *If Anyone Builds It, Everyone Dies*, 144-149.

99. Yudkowsky and Soares, *If Anyone Builds It, Everyone Dies*, 97–98.

100. Arthur C. Clarke, *Profiles of the Future: An Inquiry into the Limits of the Possible* (Harper & Row, 1973).

101. U.S.-China Economic and Security Review Commission, "Comprehensive List of The Commission's 2024 Recommendations," https://www.uscc.gov/sites/default/files/2024-11/2024_Comprehensive_List_of_Recommendations.pdf.

102. Sam Bowman, "The Checklist: What Succeeding at AI Safety Will Involve," *sleepinyourhat.github.io*, https://sleepinyourhat.github.io/checklist/; Dario Amodei, "The Adolescence of Technology."

103. Peter Barnett and Aaron Scher, "AI Governance to Avoid Extinction: The Strategic Landscape and Actionable Research Questions," *Machine Intelligence Research Institute*, May 2025, https://intelligence.org/wp-content/uploads/2025/05/AI-Governance-to-Avoid-Extinction.pdf; and Eliezer Yudkowsky and Nate Soares, "A Draft Treaty, with Annotations," *ifanyonebuildsit.com*, 2025, https://ifanyonebuildsit.com/treaty.

104. Tom Davidson et al., "AI-Enabled Coups: How a Small Group Could Use AI to Seize Power," *Forethought*, April 15, 2025, https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power.

105. Haydn Belfield, "Domestic Frontier AI Regulation, an IAEA for AI, an NPT for AI, and a US-led Allied Public-Private Partnership for AI: Four Institutions for Governing and Developing Frontier AI," *arXiv*, July 8, 2025, https://arxiv.org/abs/2507.06379; Aaron Scher et al., "An International Agreement to Prevent the Premature Creation of Artificial Superintelligence," *arXiv*, November 13, 2025, https://arxiv.org/abs/2511.10783; Cullen O'Keefe, "Chips for Peace: How the U.S. and Its Allies Can Lead on Safe and Beneficial AI," *Lawfare*, July 10. 2024, https://www.lawfaremedia.org/article/chips-for-peace--how-the-u.s.-and-its-allies-can-lead-on-safe-and-beneficial-ai; Yudkowsky and Soares, "A Draft Treaty on AI"; Romeo Dean, "Verification Plan," AI 2040, https://ai-2040.com/supplements/verification-plan.

106. "Evidence for Inference Scaling in AI Cyber Tasks: Increased Evaluation Budgets Reveal Higher Success Rates," *AI Security Institute*, March 5, 2026, https://www.aisi.gov.uk/blog/evidence-for-inference-scaling-in-ai-cyber-tasks-increased-evaluation-budgets-reveal-higher-success-rates.

107. Department of Commerce Bureau of Industry and Security, "Framework for Artificial Intelligence Diffusion," *Federal Register*, January 15, 2025, https://www.federalregister.gov/documents/2025/01/15/2025-00636/framework-for-artificial-intelligence-diffusion.

108. For example, see Michelle Nie et al., "CNAS Insights: The Export Control Loophole Fueling China's Chip Production," *Center for a New American Security*, December 19, 2025, https://www.cnas.org/publications/commentary/cnas-insights-the-export-control-loophole-fueling-chinas-chip-production.

109. Allison Duettmann, "Multipolar AI Is Underrated."

110. Several studies have explored a range of possible AI futures, an approach that may be particularly useful to help readers understand how different variables may shape future AI development. Bengusu Ozcan et al., "The Future of AI: Five Possible Scenarios."; Barry Pavel et al., "How Artificial General Intelligence Could Affect the Rise and Fall of Nations."; Department for Science, Innovation, & Technology, "AI 2030 Scenarios Report HTML (Annex C)," *gov.uk*, https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/ai-2030-scenarios-report-html-annex-c; Duncan Cass-Beggs and Matthew da Mota, "Special Report: AI National Security Scenarios," *Centre for International Governance Innovation*, 2026, https://www.cigionline.org/static/documents/AI_National_Security.pdf; Koichi Takahashi, "Scenarios and Branch Points to Future Machine Intelligence," *arXiv*, February 28, 2023, https://arxiv.org/abs/2302.14478.

111. The AI research group Epoch AI tracks many of these and other indices measuring AI progress. Epoch AI, https://epoch.ai/. See also Anthropic, "Measurements for understanding the pace of AI development inside frontier labs."

Center for a New American Security
1701 Pennsylvania Ave NW, Suite 700, Washington, DC 20006
T: 202.457.9400 | F: 202.457.9401 | CNAS.org | @CNASdc