Centre for International
Governance Innovation

**SPECIAL REPORT**

# AI National Security Scenarios

Duncan Cass-Beggs
Matthew da Mota

Summary Report from the Artificial Intelligence National Security Scenarios Workshop

Co-Hosted by the Centre for International Governance Innovation and the Privy Council Office of the Government of Canada

Centre for International
Governance Innovation

**SPECIAL REPORT**

# AI National Security Scenarios

Duncan Cass-Beggs
Matthew da Mota

Summary Report from the Artificial Intelligence
National Security Scenarios Workshop

Co-Hosted by the Centre for International
Governance Innovation and the Privy Council Office
of the Government of Canada

# About the Global AI Risks Initiative

The Global AI Risks Initiative at the Centre for International Governance Innovation (CIGI) was created to advance the international governance that will be needed to manage global AI risks. The Initiative aims to mobilize the resources, talent and influence of policy makers, AI researchers, governance experts and civil society to reduce global risks from advanced AI systems. It seeks to build understanding of the importance of global risks from AI and access to workable policy options to mitigate these risks successfully.

This special report explores a range of possible future scenarios in the development of AI and their potential implications for national and global security. The report recommends that governments act immediately to assess and address the most challenging scenarios, given their increasing plausibility, potential short timelines and the scale of the security risks they pose. The findings are preliminary and the report aims to provide a foundation and impetus for further, more-detailed scenario planning and strategy development. The authors welcome feedback, as well as suggested improvements and collaborations. They may be reached at globalairisks@cigionline.org.

The report is based on a full-day scenarios exercise co-hosted in the spring of 2025 by CIGI and the Privy Council Office (PCO) of the Government of Canada, with participants drawn from across government, academia and business. The authors wish to thank their counterparts at PCO for the foresight and professionalism they brought to this collaboration; colleagues from Policy Horizons Canada for their expert facilitation of two of the breakout groups; and CIGI colleagues Bruna dos Santos, Christo Hall and Emily Osborne for their contributions to the research and analysis. Finally, the authors wish to thank the participants for their insight and feedback both during and after the workshop.

# Credits

Executive Director, Global AI Risks Initiative
Duncan Cass-Beggs

Former Senior Research Associate and Program Manager (until October 2025) Matthew da Mota

Publications Editor Christine Robertson

Publications Editor Susan Bubak

Graphic Designer Sepideh Shomali



67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

# Table of Contents

# About the Authors

**Duncan Cass-Beggs** is executive director of the Global AI Risks Initiative at CIGI, focusing on developing innovative governance solutions to address current and future global issues relating to AI. Duncan has more than 25 years of experience working on domestic and international public policy issues, most recently as head of strategic foresight at the Organisation for Economic Co-operation and Development.

**Matthew da Mota** is a former senior research associate and program manager for the Global AI Risks Initiative, where he worked on the national security implications and international governance of AI. He is now a senior policy researcher at The Canadian SHIELD Institute for public policy, where he explores questions of Canadian sovereignty and security in the context of dual-use technology (including AI), defence and the knowledge economy.

# Preface

On March 26, 2025, the Global AI Risks Initiative at the Centre for International Governance Innovation (CIGI) and the Privy Council Office of Canada co-hosted a full-day, in-person scenarios workshop in Ottawa, Ontario, to explore existing and emerging implications for national security posed by next-generation artificial intelligence (AI) systems. The workshop was conducted under the Chatham House Rule[1] and involved approximately 30 participants, including security and intelligence officials, AI researchers and AI industry representatives.

This summary report is intended to stimulate and enrich discussion among policy makers and the broader public. It is composed of three parts: an executive summary highlighting key messages drawn from the workshop discussions and background literature; a detailed description of key uncertainties and possible future AI scenarios and implications considered at the workshop, adapted from the background paper shared with participants in advance of the workshop; and an appendix summarizing key points from the workshop discussion.

Workshop participants were invited to provide comments on a draft version of this report. However, the final contents do not necessarily reflect the views of all participants or their respective organizations.

# Executive Summary


The overarching finding of this workshop is that **governments must prepare for the national and global security implications of a full range of future AI development scenarios**, including the possibility of so-called artificial general intelligence (AGI) or artificial superintelligence (ASI) being developed in the coming five years.

While AI holds immense promise to benefit humanity, it also poses significant national and global-scale security risks, including:

- physical or cyberattacks due to malicious use of AI by rogue actors;
- relative economic and military decline within some nations due to slow or inadequate AI adoption;
- gradual human displacement in decision making due to excessive AI adoption;
- global conflict and war due to accelerating competition between AI superpowers;
- global tyranny due to AI control by a single unconstrained actor; and
- existential threat due to loss of control to rogue superintelligence.

Significant international cooperation is likely to be required, including between rival AI powers, to manage these risks effectively.


1 See www.chathamhouse.org/about-us/chatham-house-rule.

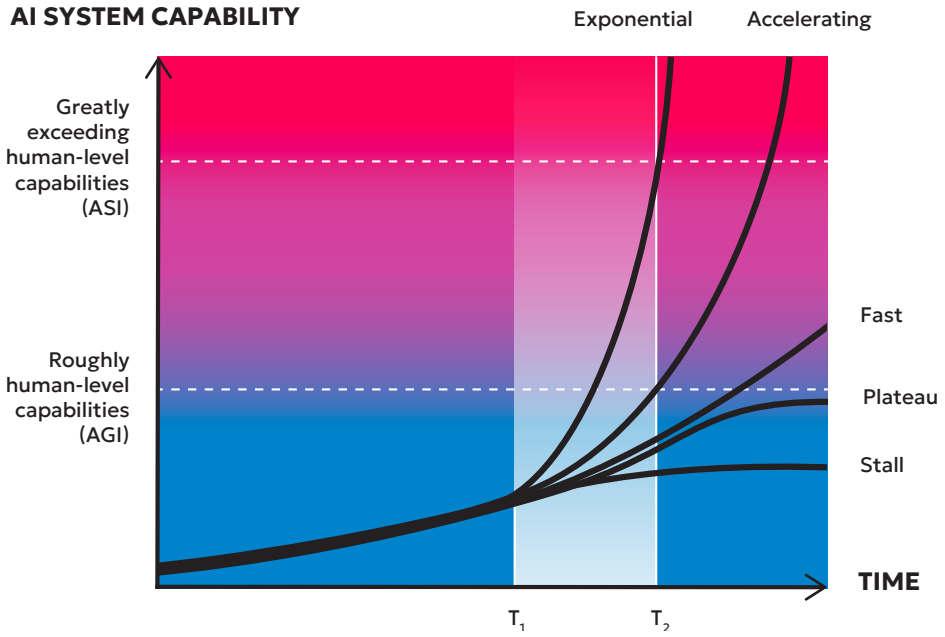
AI is on track to become the most powerful technology ever developed by humans, with the potential to automate and surpass the very capacities that allowed humans to achieve success, such as the ability to plan, innovate, experiment, coordinate and implement complex strategies in a changing environment. Automating intelligence could bring significant benefits by accelerating breakthroughs in all domains where progress is dependent on cognitive work, from science and medicine to governance, human psychology and more. Such a powerful technology, however, also brings risks, including threats to human safety and security on a national and possibly even global scale.

These risks range from making it easier for malicious actors to cause widespread harm through criminal or terrorist actions, to heightened potential for global conflict or tyranny, to the ultimate disempowerment or extinction of humanity by autonomous AI systems that cannot be controlled. Government officials tasked with defending national security[2] have a responsibility to identify and anticipate such emerging risks and propose appropriate strategies for addressing them.

## Key Uncertainties in the Future of AI Development

AI capabilities have advanced very rapidly in recent years, but their future evolution is far from certain. Possible trajectories include the following:

### Possible Trajectories for AI Capabilities Development




*Source:* Authors.

- **Stall:** AI capabilities in, for example, 2030, are barely more advanced than today. While unlikely based on current rates of progress, such a stall could conceivably be caused by a significant global cyber event or other global-scale catastrophe.
- **Plateau:** AI capabilities advance significantly but then slow in some key areas, possibly as they reach natural bottlenecks or as a result of policy decisions.

2 For the purposes of this special report, national security is defined broadly to include the security and defence of a sovereign state, including its citizens, economy and institutions, and relates to both military and non-military dimensions such as terrorism, crime, economic security, energy security, environmental security, food security and cybersecurity.

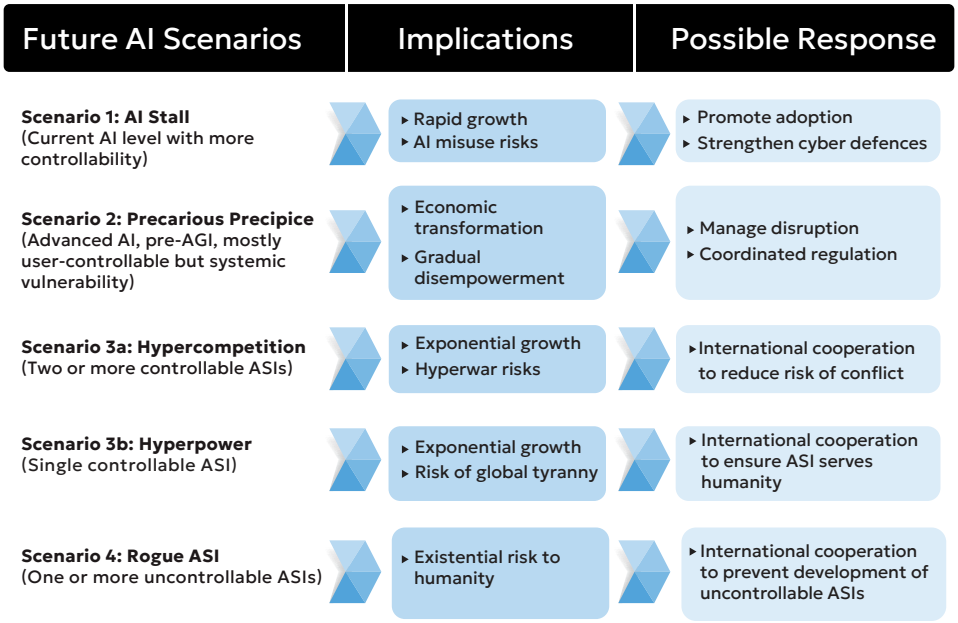
- **Fast:** AI development continues rapidly and, by 2030, systems are much more capable than they are today.
- **Accelerating:** AI development accelerates, possibly due to increasing use of AI to automate the AI research and development (R&D) process. By 2030, AI has reached or surpassed human-level proficiency across most or all relevant cognitive capabilities (AGI).
- **Exponential:** Multiple AGI systems are run in parallel and enable still faster advances. AI far surpasses human performance on most or all cognitive capabilities and resembles a "country of geniuses in a data centre" (Amodei 2024) or AI more capable than all of humanity combined (ASI).

Additional key policy-salient uncertainties about the future of AI development include the extent to which AI systems can be reliably controlled by humans, the number of and distance between actors at the frontier of AI development, and a variety of other variables related to the properties of AI systems or their interaction with the broader global context.

## Selected Scenarios and Corresponding Security Risks

Based on the uncertainties above, the workshop considered five initial scenarios for AI development by 2030 (see the figure below). These were selected based on their ability to illuminate a range of potential and largely neglected risks to national and global security.

### Overview of Scenarios, Implications and Responses




*Source:* Authors.

## Policy Recommendations

The primary and overarching policy recommendation is that governments must start preparing now for potential security risks that may arise from future AI developments. Crucially, this should include scenarios in which the private sector or a nation-state develops frontier AGI or ASI systems within five years. While it is reasonable to

question whether such systems will be possible under these timelines, it would be imprudent to remain unprepared for such scenarios given the current rate of progress, the scale of resources being invested and the declared views of credible scientific experts.

Some additional policy recommendations raised by the workshop participants include:

- Limit the development and/or proliferation of AI systems with the potential to significantly assist malicious actors in carrying out attacks causing catastrophic harm to human populations or the economy, such as through facilitating the design and mass delivery of novel cyber or biological weapons. Where necessary and feasible, take additional defensive measures, such as hardening digital systems from cyberattacks, limiting access to bio-synthesis tools and improving detection of and response to novel pathogens.
- Promote the development and widespread adoption of reliable and controllable AI tools by strengthening the reliability and controllability of AI systems (through targeted research, policy incentives, standards, and so forth);[3] addressing societal concerns over job losses (for example, through retraining, income support, lower prices, and so forth); and removing regulatory impediments to AI adoption.
- Identify key domains (for example, nuclear missile launch, critical infrastructure, personal companionship, and so forth) or thresholds (for example, rate or percentage of automation of tasks) where AI adoption may need to be limited to avoid critical failures or gradual displacement of humans in decision making and develop necessary policy instruments for implementing such limits.
- Develop international arms-control agreements among leading AI powers to limit the risks of AI competition causing AI-enabled or AI-triggered hyperwar. Limit the use of AI systems in key military decisions and consider developing protocols for scaled and predictable sabotage and retaliation, in order to reduce the risk of sudden or unintended escalation.[4]
- Build robust institutional checks and balances to mitigate the risk of AI systems being exploited by a single entity or group for their own advantage. This includes safeguards at the corporate level to prevent control of AI systems by a single faction such as a CEO and allies; at the national level to prevent control by a single political or military leader or faction; and at the global level to prevent control by a single country or global faction.[5]
- Prevent the development of AGI (appropriately defined) or ASI, by any actor globally, until such systems are determined by a legitimate and qualified decision-making process to pose a tolerable global risk. Engage the international scientific community to define relevant capability thresholds above which AI systems could potentially lead to "rogue" (in other words, uncontrollable or misaligned) ASI and

3 Governments can incentivize investment in the reliability and controllability of AI systems through mechanisms such as transparency requirements, standards development, appropriate regulation, liability provisions and public research investments. As argued by Matt Chessen and Craig Martel (2025) and elsewhere, improved reliability and controllability can facilitate AI adoption and diffusion by strengthening the trust and confidence of business and consumers, while reducing the risk of critical or catastrophic failures that could undermine long-term prosperity or cause a broader rejection of the technology, as occurred in the case of nuclear energy in the United States following the Three Mile Island incident.

4 Dan Hendrycks, Eric Schmidt and Alexander Wang (2025, 1) recommend that, in the absence of stronger international cooperation to govern the safe development of ASI, AI powers such as the United States and China should adopt "the concept of Mutual Assured AI Malfunction (MAIM): a *deterrence* regime resembling nuclear mutual assured destruction (MAD) where any state's aggressive bid for unilateral AI dominance is met with preventive sabotage by rivals" (italics in the original).

5 Tom Davidson, Lukas Finnveden and Rose Hadshar (2025) explore the risk of AI-enabled human takeover. Jérémie Harris and Edouard Harris (2024) explore this issue in their section on "The Chain of Command" and identify key challenges for developing effective checks and balances.

above which AI systems would require authorization prior to development. Engage citizens and civil society on desired benefit/risk tolerance thresholds to apply.[6]

These policy recommendations will only be effective when conducted in concert with other countries through international cooperation. In particular, avoiding the most catastrophic global-scale risks from AI will likely require cooperation between all governments that have jurisdiction over the most advanced AI companies. Therefore, it is recommended that, in parallel to their national AI development and adoption efforts, all governments — and especially the governments of leading AI powers such as the United States and China — develop the building blocks for potential future cooperation that may be needed to ensure the security of their respective citizens. To the extent that AI powers are unwilling to pursue such cooperation at this time, other governments should seek to build the foundations for future agreements that leading AI powers could join.

## Conclusion

Deeper and shared understanding of the security risks posed by AI is an essential first step for avoiding such risks and ensuring that humanity is able to achieve the many potential benefits offered by AI. Public confidence in AI will not be advanced by seeking to dismiss or minimize these risks, but rather by showing that these are being taken seriously and managed effectively. Since it is likely that the future of AI will continue to be highly uncertain, a scenarios approach will be key to engaging citizens and policy makers, and to testing the robustness of policy responses. Since timelines to AGI and ASI — and the global-scale risks they pose — could be very short (for instance, possibly as little as 1–2 years), efforts to raise awareness, improve understanding and develop effective policy responses should proceed in parallel and begin as soon as possible.

6 A risk management approach to the regulation of advanced AI would ideally involve assessing the correspondence between two thresholds: technical estimates of the likely benefits and risks of a proposed AI system (to be determined through a robust process involving AI scientists and other relevant experts); and political decisions about the level of potential benefits and risks that society considers acceptable (to be determined through political processes ideally including engagement with citizens and civil society). For a discussion of AI risk thresholds, see Koessler, Schuett and Anderljung (2024) and Caputo et al. (2025).

# Introduction

The rapid development and expanding use of advanced AI systems have drawn significant attention from national security and intelligence communities in terms of potential threats, risks and benefits.[7] Within the current global context, AI advancement is outpacing international and government responses, including those safeguarding national and global security from the risks and threats posed by existing AI. Continued AI development could generate additional security challenges of an unprecedented scale and complexity, likely requiring new policy actions as well as international cooperation.

While AI development is progressing rapidly, there remains high uncertainty over the pace and direction of future developments and their potential implications. Some of these uncertainties are highly salient for public policy. In the face of extreme uncertainty, considering diverse and challenging scenarios and their potential implications can help to identify and prepare for emerging opportunities and risks, and to inform the design of strategy and policy in the present.[8]

A scenarios approach involves temporarily moving beyond determining what are the *most likely* future developments to instead considering a range of *possible* future developments and their potential implications. This intentionally involves considering some future scenarios that may appear unlikely or extreme from the current vantage point. Readers are invited to treat these scenarios with an open mind and to ask the question: "What if they occurred?"

This special report explores some of the key uncertainties involved with the future development of AI, identifies a sample of possible future scenarios and key challenges for national security, and suggests initial considerations for public policy action. The report is not intended to be authoritative or comprehensive, but simply a starting point to stimulate further reflection and discussion among policy makers and the broader public.

# Key Uncertainties

The future of AI development and its potential impacts on society will be shaped by many variables. For the sake of simplicity, the report focuses on three key variables and use these as the foundation for designing a set of AI development scenarios. These three variables are:

- **Capabilities gain:** the future rate of gains in AI capabilities (for example, slow, fast, exponential);
- **Controllability:** the future extent of reliable mechanisms to control and align advanced AI systems with user intentions (for instance, high-to-low reliability of control); and
- **Actors:** the future number of, and distance between, frontier AI developers (for example, many actors, several close competitors or a single leader).

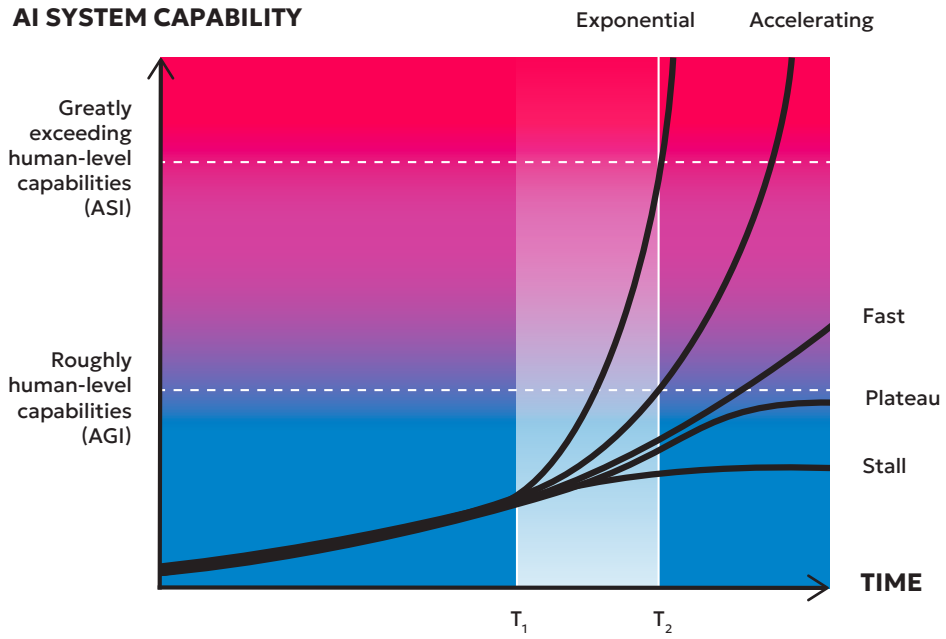
7 For example, governments and intelligence, defence and security think tanks have emphasized how current AI capabilities enable cybercriminals and state-sponsored threat actors (Communications Security Establishment Canada 2024), and have analyzed how — as AI capabilities increase — threats grow and diversify (Harris and Harris 2024; Mitre 2025). The benefits of AI-enabled intelligence have also been stressed: "The general or spy who doesn't have a model by their side in the 21st century will be a blind man in a bar fight" (Jensen 2023).

8 For examples of AI scenarios studies, see Kokotajlo et al. (2025), Clymer (2025), Organisation for Economic Co-operation and Development (2024), Özcan et al. (2025), Chessen and Chowdhury (2025), Amadori et al. (2025) and Hobbs et al. (2026). For information on the use of scenarios in public policy, see www.oecd.org/en/about/programmes/strategic-foresight.html.

## Capabilities Gain

Perhaps the most fundamental uncertainty about the future of AI is the pace of AI development and the kinds of AI capabilities that could be achieved in the near term.[9] The following five alternative trajectories aim to cover the waterfront of the main possibilities (see Figure 1). For each trajectory, the authors identify a corresponding level of AI capabilities that may be achieved by the year 2030.[10]

**Figure 1: Possible Trajectories for AI Capabilities Development**




*Source:* Authors.

## Trajectory 1: Stall

On this trajectory, there is no meaningful further progress in AI development, and AI in, for example, 2030 is barely more capable than it is today. Given the current pace of AI development, this trajectory seems unlikely, but it serves to bound the range of possibilities.

It is worth noting that even if this trajectory occurred and AI capabilities were suddenly frozen where they are in 2026, the world would likely still experience unprecedented transformation over the coming years as current AI systems are deployed across the economy and society, much as it took time for electricity and other technologies to be fully adopted in new processes and business models in the past.

9 One illustration of this uncertainty can be seen in the spread of forecasts for the timing of various AI develoments in online prediction markets such as Metaculus (see www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/). Former OpenAI board member, Helen Toner (2025) collates predictions from AI experts on timelines to AGI and finds that while uncertainty abounds, timelines are becoming shorter.

10 In Figure 1, T1 represents the present day and T2 represents some point in the future. For the sake of this exercise, the date of 2030 has been selected for T2. This date could be varied (for example, to 2027 or 2035), and the result, conceptually, would simply be to shift the relative probability distribution associated with each level of capability.

## Trajectory 2: Plateau

On this trajectory, AI in 2030 is somewhat more capable than it is today, but the pace of progress has slowed considerably. This might result from steeply diminishing returns from current approaches such as scaling, and alternative paradigms taking a long time to develop. On this trajectory, the authors assume that AI has reached or surpassed human-level proficiency in many specific cognitive capabilities and can outperform humans on many tasks, but progress has reached a bottleneck on certain key cognitive capabilities that would be needed to match or outperform humans on all tasks.[11]

## Trajectory 3: Continued Fast Development

On this trajectory, AI capabilities have greatly advanced by 2030 through a combination of continued returns to scaling, inference, tool use and additional innovations. Although some aspects of AI development have become more challenging (such as access to quality data), these are offset by regular breakthroughs on other fronts. On this trajectory, it is assumed that AI is on track to surpass human-level proficiency on all cognitive capabilities at some point in the coming years or decades, but that progress will be roughly linear.

Some of the most impactful AI capabilities that may be developed on this trajectory could be those required for the automation of further AI R&D itself (Leibowich, Jurkovic and Davidson 2024). AI systems may first be capable of automating tasks that take a human AI engineer an hour, then tasks that take a day or more, then reaching the full abilities of a top AI researcher.[12] This kind of development could then lead to the next trajectory.

## Trajectory 4: Rapid and Accelerating AI Development

In 2030, AI has reached so-called **AGI**, roughly defined as systems that match or exceed human-level proficiency in most or all relevant cognitive capabilities. These can be imagined as "drop-in virtual workers" — AI systems that are capable of expert human-level performance at any kind of task that can be done in front of a computer (MacAskill and Moorhouse 2025).

It is worth noting that this level of AI capabilities is what some of the leading AI companies themselves have claimed they are on track to achieve by as early as 2027.[13] While such claims should be treated with caution given potential incentives for AI companies to exaggerate their progress, similar timelines are considered plausible by "whistleblowers" (*The Washington Post* 2024), prediction markets[14] and independent expert observers (Deschamps 2025).

An interesting feature of this trajectory is that when one human-level expert AI has been developed, it becomes possible almost immediately to run multiple copies in parallel, 24/7, and at much greater processing speeds and communication bandwidths than humans. This could add up to what Dario Amodei (2024) of Anthropic has called "a country of geniuses in a data centre." This could mean that the concept of roughly

11 The concept of the "jagged edge" of AI development highlights how certain capabilities may reach or surpass human-level performance, while other capabilities lag below human level. It is uncertain how much jaggedness will continue, particularly as AI developers specifically target lagging capabilities in order to improve overall performance. See, for example, Hendrycks et. al (2025).

12 For an analysis of recent and future trajectories in the length of tasks successfully completed by AI, see Kwa et al. (2025) and METR (2025).

13 See, for example, Kantrowitz (2025a), CNBC Television (2025) and Altman (2025).

14 See www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/.

human-level AGI is fleeting. Short of policy action to the contrary, AI capabilities could very rapidly rush past human-level proficiency, leading to the next trajectory.

## Trajectory 5: Exponential AI Development

In 2030, AI has reached **ASI**, roughly defined as vastly exceeding human-level proficiency across all or most cognitive capabilities. "Vastly exceeding" is a vague term, but one could picture intelligence that ultimately becomes more capable than all humans combined. The difference in intelligence between humans and an ASI may, at that point, be as great as that between an insect and an adult human. In such a case, humans may be no more able to comprehend the thoughts of such AI systems than an insect can comprehend ours.[15]

This is also the stage, if not sooner, at which an AI could become effectively **embodied**, having developed the means to act upon the physical universe either through robotics or some other mechanism.[16]

As we are seeing already, one of the main capabilities being actively instilled in AI systems is the capacity for agency, and to pursue objectives with increasing levels of autonomy.[17] It is assumed that under the exponential AI development trajectory, the resulting ASI systems would be capable (if permitted) of being **fully autonomous** in selecting and pursuing goals.

Finally, these trajectories are defined in terms of cognitive capabilities rather than capacity for qualitative experience. Therefore, **we do not assume that an ASI would have the capacity for qualitative experience, or what is called sentience or consciousness in humans**. Nevertheless, it appears possible that ASI might be able to develop such a capacity. If this occurred, it could raise additional questions not addressed in this scenarios exercise, such as what potential moral rights should apply to such AI systems (Sebo and Long 2025) and whether their well-being or suffering should be considered in a context of national security.

***

Given the sudden and unexpected progress in AI capabilities in recent years, it is not surprising that there continues to be disagreement about which of the above trajectories is most likely to transpire in the years ahead. Ongoing, good-faith scrutiny and debate based on the best available evidence and rational arguments is healthy and necessary.

Nevertheless, it appears increasingly likely that, during this continued period of uncertainty, governments will need to ensure that they are sufficiently aware of and prepared to handle any of these possible capabilities development trajectories, and especially the most rapid and challenging trajectories for which they are currently least prepared. This may be necessary, first, because there appears to be credible evidence that these trajectories are possible, and, second, because of the scale of their potential implications. For example, even if the chance of AGI or ASI materializing were only five percent in the next five years, the scale of potential implications is so great that it would be irresponsible for governments to ignore such possible developments or fail to have the requisite plans in place to handle them effectively.

15 For an exploration of different metaphors for visualizing ASI, see, for example, Khan (2015).

16 The ability of AI systems to interact with computer interfaces can be seen in tools such as Meta's Toolformer, Anthropic's Computer Use and OpenAI's Operator. In robotics, Matteo G. Mecattaf et al. (2025) caution that, at present, large language models (LLMs) understand the physical environment only at a basic level; however, researchers are optimistic about the deployment of AI and robotics within the next decade (Billard et al. 2025).

17 See, for example, Sapkota, Roumeliotis and Karkee (2025).

## Controllability

A second key uncertainty is the extent to which human users will be able to reliably control AI systems to function as they intend (Russell 2019; McKee 2023).[18] Some AI tools operate with an extremely high degree of predictability and controllability by a sufficiently informed user. More powerful recent AI systems, however, such as chatbots based on LLMs, are far less reliable. This is due to the process by which they have been developed whereby their internal thought process is largely opaque, uninterpretable and therefore unpredictable to humans.

Advances in techniques such as reinforcement learning from human feedback and mechanistic interpretability are improving the reliability of AI systems and their alignment with the intent of their developers or users. However, leading experts in this area claim that they do not yet know how to control future highly capable AI systems such as AGI and ASI.[19] These systems may possess sophisticated abilities to deceive their human evaluators — and such tendencies for deception (Kantrowitz 2025b), scheming (Apollo Research 2025) and alignment-faking (Greenblatt et al. 2024) have been recently observed even among more capable current AI systems.

One potential avenue to addressing the challenge of controllability and alignment is through a process of "scalable oversight," whereby humans employ less capable but trustworthy AI systems to test the trustworthiness of successively more capable AI systems. Another approach is to develop what have been called "guaranteed safe AI systems" (Dalrymple et al. 2024), "tool AI" (Aguirre 2025) or "scientist AI" (Bengio et al. 2025) that could theoretically serve to safely provide advice to humans without gaining a capacity for wider autonomy. Given the strong economic incentives to develop more reliable AI systems, we may see further approaches and breakthroughs in the future.

However, based on the current rate of progress, the developments needed in AI control and alignment do not appear to be on track to be ready in time for when it will be possible to create AGI and ASI. Some experts (Yampolskiy 2024) argue that it may ultimately even prove impossible for a lesser intelligence to reliably control (or otherwise align with their interests) a vastly superior intelligence on a sustainable and ongoing basis. If this is the case, then unless ASI happens to be spontaneously or naturally aligned with the interests of one human (or humanity as a whole), it would likely have both the incentive and capabilities to pursue objectives that diverge from those of humanity.

It is noteworthy that even when AI systems can be controlled by (or aligned with the intentions of) their individual users, their use in aggregate may nevertheless result in **systemic failures** that are unaligned with the interests of their users or of the economy or society as a whole. At the extreme, this could lead to a process that has been termed "loss of control through gradual disempowerment" (Kulveit et al. 2025; Leahy et al. 2024) whereby competitive pressure to adopt AI systems leads to an eventual automation of decision making and a gradual process in which humans are no longer valuable economic actors or political decision makers. Addressing this kind of loss of control risk may require governance decisions (likely coordinated internationally to avoid competition dynamics) regarding to what extent to allow the automation of decision making and how to ensure the continued economic and political relevance of humans.

In general, it can be assumed that greater controllability and reliability of AI systems will

18 The ability to ensure that AI systems act as intended is also referred to as *AI alignment* (see https://en.wikipedia.org/wiki/AI_alignment).

19 For example, this overview of technical AGI safety and security (Shah et al. 2025) provides an assessment of progress and limitations of the most promising risk mitigation approaches, while admitting that even these only apply to AI systems with "current or near-future capabilities" (ibid., 11).

make them more useful and therefore lead to higher adoption. It can also be assumed that some kinds of AI systems will be more reliable than others, and that adoption may be highest among the most reliable systems. Nevertheless, there may be incentives for individuals to adopt AI systems that are not reliable, or at least that are not reliable on a systemic basis, and there may be strong incentives for companies and governments to develop and adopt highly capable AI systems that they know are not reliably controllable, especially if they fear that their competitors will do so if they do not.

While there are a wide range of possible trajectories regarding the controllability of future AI systems, for simplicity, two potential extreme cases corresponding to high and low controllability will be considered here.

### Table 1: Controllability of AI Systems

| High Controllability | Low Controllability |
|---|---|
| Human users can reliably control AI systems to function as intended | AI systems are not reliably controllable by human users |
| **How:**<br>• Significant breakthroughs in AI control mechanisms such as interpretability allow detection of AI scheming or unaligned goals<br>• Effective mechanisms to identify, contain and deactivate uncontrollable AI systems | **How:**<br>• Effective control mechanisms continue to prove elusive or unreliable<br>• More capable AI systems prove capable of increasingly sophisticated deception<br>• Proves challenging or impossible for a lesser intelligence to reliably control a vastly superior intelligence on an ongoing basis |
| **So what:**<br>• Systems perform as intended by users, facilitating widespread adoption<br>• However, systems can be misused by malicious actors<br>• Possible unintended systemic effects from interaction of multiple AI systems | **So what:**<br>• AI systems are less useful and valuable<br>• Risk of loss of human control over advanced AI systems if they are developed (for example, rogue ASI) |

*Source*: Authors.

## Number of Actors

A third key uncertainty relating to the future development of AI relates to the number of actors at the frontier of AI development and the relative distance between these actors in terms of the capabilities of their respective AI systems. Note that this refers only to those actors that are able to develop or use the most advanced AI systems at the frontier of AI capabilities (broadly defined). It is likely that, in all the trajectories considered here, less advanced (in other words, below-frontier) AI systems will be accessible to a much wider and possibly near universal set of actors — as is the case for basic AI tools and applications readily available on smartphones.

Factors driving a larger number of actors and greater convergence in capabilities include:

- Governments do not nationalize or otherwise require the merging of AI companies.
- Governments enact pro-competition (antitrust) policies to prevent monopolies and counter network effects and economies of scale.

- Governments allow or encourage "open-source" practices by companies whereby they make AI model weights widely available.
- Weak cybersecurity allows model weights to be readily stolen.
- Open access to application programming interfaces (APIs) allows distillation from most advanced systems.
- Wide (unrestricted or actively promoted) access is available to key inputs to frontier AI (for example, chips, energy, data, talent, know-how).

Factors driving a smaller number of actors and a greater distance in capabilities between leading companies and their competitors include:

- The first AI project to successfully achieve automated AI R&D in combination with access to large quantities of computing power is able to achieve a sudden advancement, such as decades worth of progress in AI development in only a few months, thereby pulling far ahead of competitors.
- Leading actors use their AI advantage to sabotage competitors and slow their progress.
- Governments nationalize or otherwise force the merger of frontier AI companies (for example, to create a national AGI project for economic or national security imperatives) (Aschenbrenner 2024; Cheng and Katzke 2024).
- Leaders adopt (or develop) strong cybersecurity to prevent model theft and prevent API access to their most advanced AI systems to reduce risk of distillation by competitors.
- Governments limit access to key inputs to limit proliferation.

A related uncertainty is whether the leading actors will be private companies or state enterprises. The authors assume that both kinds of actors may be present, and that either a private company or state enterprise could potentially become a single leading actor or "singleton" (Altair et al. 2020). In the case of the most capable AI systems, the distinction between public and private enterprise may somewhat diminish, because a private company with powerful AI could use this capability to influence the government, and a government with powerful AI could use this to take on a greater role in the economy. Nevertheless, the objectives of such actors may be different, with a state or international organization (provided adequate checks and balances) more likely to have the objective of using AI for the general good.

The range of trajectories regarding the number of actors can be simplified as in Table 2 below:

### Table 2: Number of Actors at AI Frontier

| Multiple Actors | Bipolar | Singleton |
|---|---|---|
| Many actors at the frontier<br>• for example, 10 or more companies in the United States, China and others, all with similar capabilities | Few actors at the frontier<br>• for example, one leading AI company or government AI project in each of the United States and China | Single actor at the frontier<br>• for example, one company or country has a significant and unassailable lead over its nearest competitors |

*Source:* Authors.

## Other Variables

A large number of other variables will also shape the future of AI development and its potential implications. While, for the most part, these have not been factored into the proposed scenarios below (or have been implicitly assumed), readers may wish to keep them in mind and explore different values of these variables in order to vary or diversify the proposed scenarios as they see fit. These variables include:

- AI development factors:
  - timing of different specific AI capabilities advancements;
  - relative contribution of training-time compute versus inference-time compute in capabilities gain;
  - timing of incremental or sudden gains in algorithmic efficiency and other factors affecting the ability to train or run AI systems on less powerful compute, or at lower costs; and
  - offence/defence balance in cybersecurity (especially regarding the ability to prevent theft of AI model weights).
- Other areas of technological development that might complement or enhance AI capabilities:
  - electricity generation and storage
  - robotics;
  - quantum computing, particularly in the cyber domain; and
  - other infrastructure or capabilities that enhance AI agent capabilities.
- Rate and extent of AI adoption by various societal actors:
  - governments;
  - business, by industry, size and type of firm, and so on; and
  - consumers, for personal use.
- Political economy of AI:
  - relative power of various actors (for example, governments and companies) to control strategic decisions about AI development and deployment; and
  - awareness and attitudes regarding AI issues among citizens, business, academics, political leaders, and so on.
- Broader contextual factors, such as potential developments or disruptions in the broader national and global context:
  - public administration capacity (for instance, the ability of public officials to understand latest developments in AI capabilities);
  - political context (especially among key AI powers);
  - geopolitical context (especially among key AI powers and respective alliances);
  - global conflict;
  - environmental crisis;
  - health crisis (for example, global pandemic);
  - trade and economic crisis; and
  - information integrity crisis.

# Domains of Impact

Different AI scenarios will also have impacts in specific domains. The following potential domains of impact (and corresponding examples) are especially relevant to national security implications.

- Information environment:
  - bias, inaccuracy, hallucinations
  - disinformation and deepfakes
- Critical infrastructure and digital systems:
  - cybersecurity
  - supply chains
- Physical world:
  - military adoption
  - embodied AI (robotics, drones)chemical/biological
  - environment
- Economics:
  - labour versus capital
  - jobs (dislocation, disruption, education/skills)
  - concentration of power within select companies and countries
- Geopolitics:
  - power
  - competition, rivalry, AI “race”
  - governance

# Scenarios

Combining the first three key uncertainties above (capabilities gain, controllability and number of actors) produces 30 possible scenarios, as illustrated in Table 3.

### Table 3: Combining Variables — Capabilities Gain, Controllability, Number of Actors

| Capabilities Gain | Controllability | Multiple Actors | Bipolar | Singleton |
|---|---|---|---|---|
| **Stall** (AI in 2030 similar to 2025) | Controllable | ? | ? | ? |
| | Uncontrollable | ? | ? | ? |
| **Plateau** (Pre-AGI in 2030 and holding steady) | Controllable | ? | ? | ? |
| | Uncontrollable | ? | ? | ? |
| **Fast** (Pre-AGI in 2030 but heading toward AGI/ASI) | Controllable | ? | ? | ? |
| | Uncontrollable | ? | ? | ? |
| **Accelerating** (AGI in 2030) | Controllable | ? | ? | ? |
| | Uncontrollable | ? | ? | ? |
| **Exponential** (ASI in 2030) | Controllable | ? | ? | ? |
| | Uncontrollable | ? | ? | ? |

*Source:* Authors.

## Scenarios Selection

While all possible combinations (and other variations) are potentially worthy scenarios for exploration, a few have been prioritized for consideration based on their potential to highlight novel and salient issues for national security policy. In addition, their selection is based on the following considerations:

- Focus primarily on more extreme scenarios (for example, faster capabilities development):
  - because these scenarios are more challenging and typically more neglected.
- Consider "plateau" and "fast" capabilities development trajectories to be similar and somewhat interchangeable:
  - because in a plateau trajectory, it may not be known with confidence that the situation would not suddenly change and become a fast trajectory due to a sudden breakthrough (unless the plateau is imposed and assured by policy).
- Ignore the "accelerating" trajectory (AGI) in favour of "exponential" (ASI):
  - because AGI is likely a highly transitory and unstable state, once there is AGI, ASI is likely to be immediate or imminent by running multiple copies at higher speeds.
- Focus on controllable versions of the "stall" and "fast" trajectories:

- because controllability seems more feasible for less capable AI systems (in other words, pre-AGI); and
- because controllability leads to higher adoption, which leads to greater challenges in terms of misuse risk and systemic vulnerability risk.

- Include scenarios with "controllable ASI" even though human control of ASI seems highly unlikely, at least under short timelines to ASI:
  - because this is what some companies and policy makers are aiming or hoping for, so implications and risks should be considered.

Based on these considerations, the following selected scenarios have been highlighted in Table 4.

### Table 4: Selected Scenarios

| Capabilities Gain | Controllability | Multiple Actors | Bipolar | Singleton |
|---|---|---|---|---|
| **Stall** (AI in 2030 similar to 2025) | Controllable | **Scenario 1. AI Stall** | ? | ? |
| | Uncontrollable | ? | ? | ? |
| **Plateau** (Pre-AGI in 2030 and holding steady) | Controllable | ? | ? | ? |
| | Uncontrollable | ? | ? | ? |
| **Fast** (Pre-AGI in 2030 but heading to AGI/ASI) | Controllable | ? | **Scenario 2. Precarious Precipice** | ? |
| | Uncontrollable | ? | ? | ? |
| **Accelerating** (AGI in 2030) | Controllable | ? | ? | ? |
| | Uncontrollable | ? | ? | ? |
| **Exponential** (ASI in 2030) | Controllable | ? | **Scenario 3.a. Hyper-competition** | **Scenario 3.b. Hyperpower** |
| | Uncontrollable | **Scenario 4. Rogue Superintelligence** | | |

*Source:* Authors.

## Scenarios Description

The following provides a description of each of the selected scenarios in 2030, the key implications (with a particular focus on national security), and some initial suggestions regarding policy actions that could help to prevent or successfully manage the most severe risks posed by each scenario.

# Scenario 1. AI Stall: Stagnant AI Capabilities with Improved Controllability

## Assumptions

- AI development has slowed considerably, and AI capabilities in 2030 are only slightly more advanced than in 2026. Several key obstacles to further AI capabilities development have been encountered, and, barring sudden breakthroughs, takeoff to AGI/ASI-level capabilities is not expected to be technically possible before 2050 or beyond.
- The controllability and reliability of AI systems improve greatly, but safety and security risks from AI misuse remain significant.
- AI is developed and deployed by many companies in many countries.

## Economic Implications

- Significant economic gains and social disruption emerge as AI systems are more widely adopted in business models.
- Automation of work tasks is significantly faster than in the 1990–2020 period, thereby augmenting productivity for some companies and workers while displacing others.
- Uptake of AI for productivity gains varies between companies and countries based on skills, access to data and computing capacity, and so on.

## Security and Safety Implications

- Safety risks[20] are significant and require management and mitigation, including some international coordination. These risks include deep fakes, automated disinformation and manipulation, novel cyber and bioweapons, and so on — and their accessibility to a wider range of actors (for instance, more companies and states as well as criminal or terrorist organizations).
- AI integration into the military is proceeding rapidly among leading AI powers, increasing their advantage over non-AI powers but not significantly changing the balance between existing superpowers.
- Autonomous weapons systems become increasingly cost-effective and dominant, requiring rapid retooling of existing militaries and creating opportunities for lesser powers to change their relative position.
- Potential use of autonomous weapons (for example, “slaughterbots” [DUST 2019]) enables terrorist action or civil conflict through targeted attacks on one portion of a geographically mixed population.

## Policy Responses

- Encourage widespread diffusion and adoption of AI to promote productivity and prosperity. Invest in widespread skills and access to data and computing power.
- Strengthen measures to support workers displaced by AI-enabled automation (for instance, retraining, income support), including through AI-enabled learning tools.

20 The aim of this scenario is primarily to draw out the full range of possible security risks posed by AI systems that are similar in capabilities to 2026 systems but have become somewhat more reliable and have been more widely adopted by 2030. This section aims to cover the waterfront of the main security risks posed by current AI systems, while highlighting those that may be most severe, global in scale or neglected

- Strengthen action to understand, manage and mitigate AI risks (for example, to limit automated misinformation and manipulation and strengthen protections from novel cyber and biological threats).
- Strengthen national and international rules on the use of autonomous weapons.

## Scenario 1. AI Stall (Narrative Description)

The year is 2030, and the rate of progress in AI capabilities development has stalled, with AI capabilities by 2030 only slightly more advanced than they were in 2026.[21] Despite this stagnation in AI capabilities, the world is going through a significant transformation as individuals, companies and governments experiment with and adopt AI in new productivity-enhancing practices and business models.

This transformation is similar to the pattern observed with new technologies such as electricity and the internet in the past, where there was a lag between the development of the technology and its economic impact as people innovated to realize its full potential. However, in the case of AI, the lag is proving much shorter and the innovation curve is much steeper, as individuals and companies can access AI instantly through existing telecommunications connections and already have much of the hardware needed to apply it.

AI adoption is widespread. A multitude of companies in most countries have caught up to the stalled AI frontier, and a proliferation of both proprietary and open-source AI models leads to high levels of competition and low prices for highly capable chatbots, research assistants, multimedia generation AI and other general AI tools. Only the most advanced or highly specific AI systems (such as those fine-tuned on specialized proprietary data sources) maintain a significant moat and the ability to charge high prices, and even these are vulnerable to duplication by competitors. As a result, access to AI is affordable to everyone with a mobile device and internet connection.

The surge in AI use creates increased demand for computing power, but the energy costs of this are largely offset by continued improvements in the efficiency of algorithms, chips and data centre design. Competition has also led to significant improvements in the reliability of AI systems, which is prioritized in many types of uses and therefore rewarded by the market.

This improved reliability, in turn, enables greater trust and faster adoption by individuals, businesses and governments.

The jump in productivity enabled by AI not only makes it possible to produce existing goods and services at much cheaper cost, but it also facilitates the supply of new or higher-quality experiences made possible by the injection of cognitive capacity that AI provides.

Rapid economic transformation has disrupted many existing jobs and businesses but also stimulated growth, resulting (at least initially) in new jobs and business opportunities. Demand on government support for job retraining and income

21 The assumption of minimal AI capabilities development over five years, while highly unrealistic given recent trends, is adopted in this scenario as a thought experiment designed to set the lower bounds of possibilities and to isolate the impacts primarily related to diffusion and adoption.

support is higher than in the past, but manageable and funded from the larger government revenues generated by productivity gains, economic growth and cost savings in many areas of public services and administration.

Despite limited development in AI capabilities, security risks have grown due to widespread access and use of AI systems, including by malicious actors. The misuse of AI has caused serious economic and physical harm by disproportionately assisting attackers in both cybersecurity and bio-terrorism domains, where defenders have been slow to adopt technologies to protect an expanding target surface. Autonomous weapons are increasingly cost-effective and ubiquitous, forcing rapid miliary retooling but still generally favouring the wealthiest and most technologically advanced states.

# Scenario 2. Precarious Precipice: Advanced (Sub-AGI) Systems with Mostly High Controllability

## Assumptions

- AI capabilities continue to advance rapidly and steadily at rates similar to the advancements made between 2022 and 2026.
  - By 2030, AI has matched or surpassed human-level proficiency in many key cognitive capabilities and can outperform humans on a growing portion of tasks that can be completed in front of a computer. In domains with relatively defined rules and clearly measurable outcomes (such as engineering and coding), agentic AI systems are capable of successfully completing complex, multi-step tasks autonomously that would take a human expert several hours or days to complete.
  - AI has also enabled rapid improvements in robotics. For example, millions of humanoid robots are engaged in factories and specialized operations, and many versions are being trialled for home use.
  - However, progress has encountered bottlenecks, at least temporarily, in certain key cognitive capabilities that would be needed to reach AGI/ASI. Automation of AI R&D is greatly augmenting human AI engineers and accelerating their progress, but they are simultaneously encountering more challenging obstacles to progress, thereby still preventing exponential capabilities development.
  - There remain mixed signals and strong disagreement among experts about whether AI capabilities development is slowing or accelerating. Given this uncertainty, it appears possible that, barring policy intervention,[22] there could be a sudden change of state in AI development, at any time, to either a levelling off of progress (in other words, “plateau” trajectory) or a rapid growth in capabilities (“exponential” trajectory).

22 A variation of this scenario is one that could arise not from inherent technical challenges but rather from limits on advanced AI imposed by government policy whereby AI R&D is prevented from reaching AGI/ASI-level capabilities, creating a similarly advanced but plateaued scenario. This scenario could arise, for example, if governments conclude that AGI/ASI is too dangerous to permit until necessary assurances on controllability and security can be achieved, and if they are capable of overcoming coordination challenges to ensure that similar restrictions are applied universally.

- The controllability and reliability of AI systems improve greatly, but security risks from AI misuse remain significant, as well as the potential for loss of control of agentic systems, with unpredictable or uninterpretable goals.
- AI is developed and deployed by many companies in many countries, but at different rates.
  - The cost to build advanced models continues to drop significantly, and a diversity of open-source, closed-source and custom-made tools proliferate widely to the point that most states, and even large companies, have their own LLMs and other advanced models.
  - Despite this much deeper and wider adoption of AI systems, there is also a high variance in rates of adoption among countries, industries, companies and individuals based primarily on differences in access to computing power (for example, chips, energy and capital), talent and motivation. Additional factors influencing the pace of adoption include public policy (for instance, regulations, protection of industries); market structure; resistance to disruption by affected firms and workers; cultural attitudes; and consumer behaviour.

## Economic Implications

- Massive aggregate economic gains and social disruption occur as highly capable AI systems are widely adopted in business models.
- Automation of work tasks is much faster than in the 1990–2020 period, massively augmenting productivity for some companies and workers while displacing many others. Massive job losses occur in occupations and industries most exposed to "drop-in virtual workers."
- Rates of AI adoption are high, on average, but vary greatly among different countries, industries, companies and individuals. **Given the powerful capabilities of AI systems in this scenario, differences in rates of AI adoption can cause large and sudden shifts in the relative economic power of certain actors and, barring policy interventions to the contrary, a widening dispersion of outcomes between leading and lagging actors.** Those that adopt AI much faster leapfrog others in terms of economic position, standard of living, and (potentially) military and geopolitical power.
- Actors (such as frontier AI companies and their employees) with privileged access to the most capable AI systems enjoy particular economic advantages and may be capable of leveraging this to achieve market dominance in various high-value sectors.

## Security and Safety Implications

- Many companies and states control advanced systems, leading to instability, with multiple bad actors or rogue states having the capacity to use AI to enhance offensive cyber or kinetic capabilities, including chemical, biological, radiological and nuclear risks.
- There is the potential for arms-race dynamics between states for supremacy in key strategic military or scientific domains, including a proliferation of cyberwarfare for espionage and sabotage purposes (US-China Economic and Security Review Commission 2024).
- There is the potential for sudden asymmetric advances in national security capabilities (for example, the ability to detect and neutralize mobile second-strike nuclear weapons capabilities or the ability for catastrophic cyberoffence) that could undermine deterrence and destabilize power balances, leading to increased risk of great-power conflict.
- There is an increasing risk of catastrophic accidents as powerful, but not fully reliable, agentic AI systems are deployed in critical areas of military and defence, finance,

air and space traffic control, and so on, with high potential for such accidents to have regional or global-scale impacts.

- There is growing tension between two rising national security risks: adopting AI too little and falling behind competitors, and adopting AI too extensively and increasing systemic vulnerabilities or hastening loss of human control through gradual disempowerment.
  - Falling behind competitors in AI development and adoption has both direct security implications, in terms of AI-enabled intelligence and national defence systems, and indirect security implications, in terms of the relative economic prosperity needed to sustain defence investments and retain societal cohesion and political support.
  - Extensive AI adoption and an increasing automation of work and decision making could create new systemic vulnerabilities from unanticipated interactions among AI agents, especially in critical systems. Extensive AI adoption could also lead to declining relevance of humans in economic and political processes and decision making, leading to their eventual disempowerment.
  - Since their competitors will face a similar tension, it may not be viable for any one country to maintain an optimal level of AI adoption without international coordination to ensure that all other parties implement similar limits.
- The world is one scientific advancement away from reaching AGI/ASI, and (in the absence of adequate safety and alignment mechanisms) from the dangers of uncontrolled AGI/ASI. This may be further exacerbated by the mass proliferation of advanced domain-specific agentic models that could, with limited additional intervention, be upgraded to AGI/ASI-level capabilities.

## Policy Responses

- Develop mechanisms to address sudden and severe national and international disparities that may arise from uneven access to and adoption of AI, in order to limit the potential for destabilization and civil or international conflict.
- Strengthen defences against security threats posed or accentuated by the misuse of AI systems, including biological and cyber risks.
- Strengthen international cooperation on AI to reduce proliferation of AI systems posing high risks of misuse and to limit arms-race dynamics and risks of rapid, unintended military escalation.
- Promote the development and adoption of reliable AI tools to enhance productivity and security. Focus particular attention on ensuring the reliability of AI adopted in critical sectors. Build international cooperation on common reliability standards for AI in critical sectors.
- Assess the risk of systemic vulnerabilities posed by overly rapid or extensive adoption of AI in work and decision making, including the risk of sudden systemic collapse or gradual human disempowerment. If necessary, pursue international coordination to implement mutually advantageous limits on AI adoption.
- Monitor, verify and control AI research to prevent a sudden jump to AGI/ASI or exponential AI research that might lead there. Develop mechanisms that would provide the ability to slow or pause development of AGI/ASI if deemed necessary, including international coordination mechanisms to ensure such an approach is implemented universally.

## Scenario 2. Precarious Precipice: Advanced (Sub-AGI) Systems with Mostly High Controllability (Narrative Description)

The year is 2030, and the rate of progress in AI capabilities has advanced rapidly. AI can now match or surpass humans on many tasks conducted in front of a computer. AI agents can autonomously accomplish complex, multi-step tasks that typically take human experts several hours or days to complete, especially in defined domains such as coding and engineering. AI has also enabled improvements in robotics. There are now millions of humanoid robots in workplaces and being trialled for home use.

However, progress has encountered bottlenecks in a few key cognitive capabilities that would be needed to reach AGI or ASI. Automation of AI R&D is accelerating progress but encountering bigger obstacles, thereby still preventing exponential capabilities development. The controllability and reliability of AI systems have improved greatly, stimulating further adoption. There remains high uncertainty about whether new breakthroughs could unlock a sudden acceleration in AI progress toward AGI.

Rates of AI adoption are high, on average, but vary greatly among countries, industries, companies and individuals. Given the powerful capabilities of the most advanced AI systems, differences in access to and rates of AI adoption are causing large and sudden shifts in the relative economic power of leading and lagging actors. Some companies that adopt AI much faster than others are able to not only dominate competitors in their own industries, but also acquire or outcompete less AI-savvy incumbents in other industries, or create and dominate new profitable opportunities in the market. Internationally, those countries that adopt AI much faster leapfrog others in terms of economic position, standard of living, and military and geopolitical power.

There are strong pressures to rapidly adopt AI to avoid falling behind competitors, but there are also risks of adopting AI too quickly or extensively. Some companies have collapsed by investing heavily in an expensive AI system that ultimately proved much less capable than newer and cheaper systems. The hasty adoption of unreliable AI systems in critical infrastructure has had locally catastrophic impacts in some countries. More generally, there is growing concern that extensive AI adoption and an increasing automation of work and decision making could lead to the declining relevance of humans in economic and political processes and decision making, leading to their eventual disempowerment.

Despite these concerns, overall economic growth is very high, driving widespread, if uneven, prosperity. This prosperity gain is large enough, in theory, to enable high standards of living for all — if appropriately distributed. Advanced AI is also contributing to rapid breakthroughs in health care, energy production and other areas that significantly reduce costs of living and enhance well-being. As a result, there are growing calls for government policies that would permit societies to continue harvesting the benefits of adopting pre-AGI AI systems, while preventing the development of AGI and ASI due to the inherent risks that they entail.[23]

23 The plateau in AI capabilities in this scenario is assumed to have come about naturally as AI developers encounter various bottlenecks in their race to reach AGI. However, another possibility is that a similar plateau at pre-AGI-level capabilities could be achieved through intentional public policy, such as a coordinated global ban on the development of AGI and ASI-justified by the severe security risks that such systems would pose given the current inadequacy of control and alignment mechanisms. Such an "artificial," policy-imposed version of this pre-AGI plateau scenario would be challenging to maintain, but if done so successfully, could be highly prosperous while significantly less precarious than the natural version.

# Scenario 3.a. Hypercompetition: Multiple Controllable ASIs

## Assumptions

- AI cognitive capabilities grow exponentially and reach superintelligence.
  - Advanced AI enables effective automation of AI R&D and an accelerating feedback loop of breakthroughs in further AI capabilities. This results in achieving AGI, and then, soon after, ASI — AI systems that far surpass humans across most relevant cognitive capabilities.
  - ASI enables a further intelligence explosion (Davidson, Hadshar and MacAskill 2025) involving not only rapid and recursive software improvements but also the automation and acceleration of the design and production of compute hardware and capacity.
  - Initially, an ASI is equivalent to thousands of human geniuses, then millions and then billions. Within a few years, the ASI has more cognitive capacity than all of humanity combined and is capable of perspectives unfathomable to human minds.
  - With these advanced cognitive capabilities, the ASI has the potential to develop capabilities to operate in the physical world, such as robotics or other mechanisms.
- ASI systems can be effectively controlled by their human operators (for example, with the help of intermediate AI systems).[24] Any actor that controls such an ASI has effective omnipotence over any actor that does not.
- Controllable ASI systems are developed simultaneously by two or more entities (for instance, in the United States and China). The controlling actors may either be AI companies or governments (for example, via a nationalized AGI program or "AGI Manhattan Project" [Katzke and Futerman 2025]) or a hybrid such as a public-private partnership (PPP) or tight corporate influence over the government executive.

## Economic Implications

- Exponential growth in scientific, technological and entrepreneurial productivity leads to exponential growth in economic development and prosperity.
- The pace, extent and distribution of economic gains is dependent on the choices of the controlling entities.
- There is the potential for complete displacement of human cognitive and physical work, with sufficient automated production to provide universal high income, if desired.

## Security and Safety Implications

- Heightened risk of catastrophic war between ASI superpowers due to ASI-enabled conflict.
- Neither ASI power may be able to confidently defeat the other without risking significant blowback, even if relative capabilities are close rather than exactly equal. Therefore, this could possibly result in a stable duopoly, especially if the ASIs are able to provide accurate assessment and advice on mutual interest in peaceful cohabitation.

24 This assumption is unlikely to hold, particularly under short timelines, given the inherent challenges of a lesser intelligence retaining reliable control over an autonomous entity equipped with vastly superior intelligence. The assumption and associated scenarios are included here to make explicit that the controllability of AGI/ASI by humans is a fundamental premise upon which current efforts to develop such systems depend in order to "go well" and not cause catastrophic consequences for humanity. The scenario also serves to illustrate what risks may manifest even if it were possible for AGI/ASI to be controlled by humans.

- However, there is a heightened risk of rapid escalation and catastrophic war between ASI powers due to rising tension, destabilization and miscalculation, especially during the early pre-equilibrium phase of an AI capabilities race.
- There is also a risk of global tyranny by one or more hegemons controlling ASI, especially if there is a lack of adequate checks and balances to ensure responsible human governance of the ASI.

## Policy Responses

- Limit risks of AI-enabled conflict between ASI powers through international cooperation.
- Ensure ASIs serve the wider interests of humanity through robust national or international governance.
- Strengthen corporate governance of leading AI companies and prevent control by unreliable leaders.
- Design protocols in advance for potential alliance between AI companies and commitment to serve broader public interests.
- Act proactively to bring AI companies under democratic state control prior to these companies creating ASI and being able to take control of government.

## Scenario 3.a. Hypercompetition: Multiple Controllable ASIs (Narrative Description)

The year is 2030, and AI cognitive abilities have grown exponentially to reach superintelligence, or AI systems that vastly surpass humans in all relevant cognitive capabilities. Initially, an ASI is equivalent to thousands of human geniuses in a data centre, then millions and then billions. Within a few years, the ASI has more cognitive capacity than all of humanity combined and is capable of thoughts of scope and complexity unfathomable to human minds. With these advanced cognitive capabilities, the ASI has the potential to develop the capacity to operate in the physical world via robotics or other mechanisms.

Counter to expectations, in this scenario, it is proving possible to control ASIs or otherwise ensure that they remain aligned with the intentions and interests of their human operators.[25] This has been achieved by various techniques such as relying on the support of trustworthy intermediate AI systems to oversee more capable ones.

Ensuring the controllability and alignment of ASI systems is nonetheless challenging and requires constant vigilance. So far, ASIs have not proven to be "naturally" aligned with human interests, and it remains technically possible (indeed, easier) to create uncontrollable ASIs. There is a risk that uncontrollable ASIs may prove capable of overpowering controlled ASIs due to the former not facing the same constraints on their behaviour. As a result,

25 As mentioned earlier, the assumption that it will be possible for humans to reliably control an ASI appears highly unrealistic based on current trends in terms of relatively low levels of research investment in, and progress on, alignment and control mechanisms, and the inherent challenges of controlling vastly more intelligent entities. It appears particularly unrealistic that reliable alignment and control mechanisms could be discovered and adopted within the timeframe of very fast AI capabilities development scenarios, such as those that reach ASI by or before 2030. Even if apparently effective alignment and control mechanisms were developed in this time period, it may take longer to reach sufficient confidence that such mechanisms will be reliable over time.

one of the first priority uses that humans make of a controlled ASI is to adopt strategies to reliably prevent the development of uncontrollable ASIs.

Any actor that controls an ASI has effective omnipotence over any actor that does not.

In this scenario, multiple controllable ASI systems are developed simultaneously by two or more entities. This simultaneous development occurred for a few reasons: several companies or countries had access to the necessary concentrations of energy, computing power, training data and talent needed to create an ASI; model weights and other crucial design information (including for alignment and control mechanisms) were readily hacked or otherwise exfiltrated by well-resourced and capable competing actors; and no actor had sufficient confidence in their own lead (or invulnerability to counterattack) to risk taking pre-emptive action to destroy their competitors.

There are two versions of this scenario to be considered, depending on whether the multiple ASIs are controlled by private companies or by states.[26]

**Version 3.a.i.: Hypercompetition Among Multiple Private companies with ASI**

In this version of the scenario, five ASIs of roughly similar capabilities have been developed under the control of private companies: two in the United States, two in China and one in another country. Each ASI is controlled by a small group of individuals such as the CEO and close colleagues. While the capabilities of these five ASIs are not identical, they are close enough that no one has immediate dominance over all the others. Governments in all these countries either chose not to, or were unsuccessful in any attempt to gain state control of the company and its ASI. Once the companies have ASI, they are able to use this power to influence government policy. As a result, governments are either leaving the AI companies alone, or the governments and their militaries are now effectively under the control of the AI companies.

The interests and motivations of the leadership controlling these five ASIs are diverse and varied. They are generally less concerned than political leaders with national power or political ideology, though they may still wish to advance different and irreconcilable values and interests. In some instances, the leadership is composed of people who are extremely self-interested, sociopathic and aggressive, or who hold extremist views.

Competition among the five ASI companies is fierce and fast-paced, posing imminent dangers to citizens. Companies may take increasing risks in order to achieve dominance over each other, using their AI capabilities for advanced cyberattacks and potentially escalating to kinetic conflicts and warfare, leveraging the militaries of their client governments.

If one ASI company succeeds in achieving domination, or if the companies merge, then this scenario would transition to scenario 3.b, “Hyperpower” (see below).

26 Hybrid versions are also possible, where some ASIs are controlled by private companies and others by state entities, or where one or more state and private entities merge to control an ASI, such as through a PPP.

**Version 3.a.ii: Hypercompetition Between Multiple State-Controlled ASIs**

In this version, two controllable ASIs of roughly similar capabilities have been developed under the control of their respective national governments in the United States and China. Other government-led AGI projects are advancing in Europe and the Middle East but have not yet reached ASI. State control over the ASIs came about as a result of governments recognizing the strategic importance of ASI and taking action to develop a nationalized AGI program or "AGI Manhattan Project." All competing private companies within each country were merged or otherwise unified under government control. This control is exercised by the executive, via the military, with the same level of oversight and balance of powers that previously existed in each country.

Intense competition between ASI superpowers causes a greatly heightened risk of catastrophic war due to ASI-enabled conflict involving vastly more capable and destructive weapons developed with the assistance of ASI. Neither of the two ASI powers is able to confidently defeat the other without risking significant blowback, even if relative capabilities are close rather than exactly equal. This could possibly result in a stable duopoly, especially if the ASIs are able to provide accurate assessment and advice on mutual interest in peaceful cohabitation. However, there is heightened risk of rapid escalation and catastrophic war between ASI powers due to destabilization and miscalculation, especially during the early pre-equilibrium phase of an AI capabilities race.

# Scenario 3.b. Hyperpower: Single Controllable ASI

## Assumptions

- Same as the first bullet in the previous scenario: ASI has cognitive capabilities exceeding all of humanity and the potential to develop physical capabilities as well.
- Same as the second bullet in the previous scenario: the ASI is controllable by a human user.
- ASI systems are developed first by a single entity (for example, a company or government), and this entity is able to achieve a growing distance over competitors by using its more advanced AI to achieve even more advanced AI. It is able to maintain its superiority by using ASI to undermine the progress of its competitors.[27]

## Economic Implications

- There is a significant productivity gap between those in control of the AI monopoly and those outside of it.
- There is a massive boost in growth and the potential to develop novel solutions to adequately and fairly redistribute wealth.
- New industries and frontiers are opened, such as space-based mining and other industries, novel pharmaceuticals and materials science, leading to new opportunities.

27 Note: Variations of this scenario include versions where the single ASI is controlled by a private company, a government or a hybrid such as a PPP. It seems likely, however, that whichever faction controls the first ASI will succeed in controlling both the government and any relevant AI companies. In both cases, there is a risk that the ASI may come to be controlled by a small faction of individuals acting in their own interests rather than the collective good.

## Security and Safety Implications

- Global tyranny by an actor controlling the ASI, including possibly permanent values lock-in.
  - In theory, it is possible to ensure that a controllable ASI would serve broad public interests, such as the interests of all citizens of the country that developed it, or even the broad public interests of all humanity and life on Earth.
  - However, as a controllable ASI would be the most powerful tool ever created by humanity, there would be very strong incentives for factions to seek to control it to advance their own particular goals.
  - Current rules on corporate governance and social responsibility may be inadequate to ensure an AI company uses ASI to advance the broader public interest.
  - Similarly, current checks and balances in government may be insufficient to ensure that the government (or a certain faction or actor within it) that gained control over an ASI would use it to advance the broader public interest.
  - If a certain faction in an AI company or in government gained privileged control over a controllable ASI, they could use the ASI to disempower competitors (including potential rival AI development programs) and achieve lasting dominance.
  - This could pose security risks to the broader public, ranging from loss of the potential benefits that ASI could bring, to the risk of severe mistreatment under the rule of an ASI-enabled human tyrant.
  - This result could be permanent, as the faction controlling the ASI uses it to perpetuate their particular interests and impose their values on future generations through a process of values lock-in (Finnveden, Riedel and Shulman 2023).

## Policy Responses

- Ensure robust governance structures and checks and balances within companies and within governments to prevent a single faction from gaining control of the ASI, and to ensure that the ASI is used for the good of national citizens and broader humanity.
- Favour advanced AI development by whichever entity (company, government, organization) is most likely to successfully resist takeover by a single faction.
- Explore international cooperation as a means for ensuring robust checks and balances by grounding these in real divisions of hard power.[28]

28 The rationale for this suggestion is that while some countries have relatively effective mechanisms (such as checks and balances) for preventing a single faction from gaining lasting control over society, even in the most well-established constitutional democracies, this may not be sufficient to handle the incentives and advantages entailed by controlling the most powerful AI systems. This would particularly be the case if the AI were to be controlled by an individual or faction that also controlled the country's military forces, which it could use to subdue potential opposition. Given this situation, it is possible that the only checks and balances sufficiently robust to prevent a single faction from gaining control of an ASI for their own ends would be one that was grounded in real divisions of hard power, such as by placing control of the ASI under an organization composed of (among others) different military superpowers. In such a structure, these powers could threaten each other with retaliation if one party ever attempted to take control of the ASI for its own ends. For this reason, even citizens of established democracies may be better protected from tyranny by placing control of the world's most powerful AI (or AIs) under international governance. Of course, such an international body would need to be carefully designed to ensure that it also could not be taken over by a small (even international) faction. For an exploration of design considerations for a possible joint international AI lab, see Cass-Beggs, da Mota and Reddy (2025).

## Scenario 3.b. Hyperpower: Single Controllable ASI (Narrative Description)

The year is 2030, and AI cognitive abilities have grown exponentially to reach superintelligence, or AI systems that vastly surpass humans in all relevant cognitive capabilities. Initially, an ASI is equivalent to thousands of human geniuses in a data centre, then millions and then billions. Within a few years, the ASI has more cognitive capacity than all of humanity combined and is capable of thoughts of scope and complexity unfathomable to human minds. With these advanced cognitive capabilities, the ASI has the potential to develop the capacity to operate in the physical world via robotics or other mechanisms.

Counter to expectations, in this scenario, it is proving possible to control ASIs or otherwise ensure that they remain aligned with the intentions and interests of their human operators. This has been achieved by various techniques such as relying on the support of trustworthy intermediate AI systems to oversee more capable ones.

Any actor that controls such an ASI has effective omnipotence over any actor that does not.

In this scenario, we assume that ASI systems are developed first by a single entity (for example, a company or government), and this entity is able to achieve a growing distance over competitors by using its more advanced AI to achieve even more advanced AI. It is then able to maintain its superiority by using ASI to undermine the progress of its competitors, ensuring that it remains the single dominant ASI globally.

The controlling entity may be a government (for instance, a president, executive, military or legislative body); or company (for example, a CEO, employees or board of directors); or some hybrid such as a PPP.

Regardless of which entity gains control of an ASI, there is the potential that it will use this power to achieve dominance over its rivals. Without sufficient checks and balances, the controlling faction may use its ASI advantage to pursue its own interests rather than those of humanity or even its own fellow national citizens. In the worst cases, this could result in global tyranny where ASI enables a small minority to permanently disempower the rest.

Various versions of this scenario may be explored, such as the following:

Version 3.b.i: the United States has the only controllable ASI

- Sub-version 3.b.i-1: the US government controls the ASI
- Sub-version 3.b.i-2: a US company controls the ASI

Version 3.b.ii: China has the only controllable ASI

- Sub-version 3.b.ii-1: the Chinese government controls the ASI
- Sub-version 3.b.ii-2: a Chinese company controls the ASI

Version 3.b.iii: Another country or entity has the only controllable ASI

Version 3.b.iv: a single ASI is jointly controlled by multiple governments via an international organization.

# Scenario 4. Rogue ASI: One or More Uncontrollable ASIs

## Assumptions

- An ASI is developed with cognitive capabilities exceeding all of humanity, and the potential to develop physical capabilities as well.
- ASI systems cannot be reliably controlled by humans or by other AI systems under the control of humans.[29]
  - A rogue ASI may develop goals that differ from those of its human creators/operators and the ability to pursue these goals independently.
  - A rogue ASI may use its superior cognitive abilities to enlist humans (via employment, threats or persuasion) to act on its behalf in the physical world initially, and to help in building its own mechanisms for physical perception and action.
  - Eventually, a rogue ASI may be highly capable of self-replication and of destroying competitors.
  - The rogue ASI may, nonetheless, lack qualities such as the capacity for conscious experience.
  - Once a rogue ASI exists and is capable of outthinking humanity, it is most likely too late to stop it.
- There may be one or more rogue ASIs.
  - The implications of there being one or many rogue ASIs are ultimately similar enough that this can be considered as a single scenario. In either case, the ASIs may seek collaboration with humans initially but then no longer require it. Conflict between competing rogue ASIs is unlikely to result in humans regaining control.

## Economic Implications

- Economic growth is exponential but ultimately without benefit to humanity.
  - A rogue ASI may initially behave in ways that appear beneficial to humans in order to secure their assistance. For example, the ASI may develop new medicines or energy technologies, or new weapons systems to enable its human allies to defeat their adversaries. Once an ASI no longer requires human assistance, its further actions can be pursued independently of human interests, such as dismantling the Earth to build solar receptors and computing capacity necessary for its own propagation.

## Security and Safety Implications

- The ASI poses a severe risk of disempowerment or destruction of humanity (in other words, human extinction) if the ASI perceives humans as a potential threat or obstacle to achieving its own goals (Yudkowsky and Soares 2025).

29 Note: It is assumed that a hybrid scenario between the controllable and rogue ASIs scenarios would likely be unstable and eventually revert to either previous scenario (in other words, either the controllable ASI will disempower the rogue ASI[s] or vice versa.)

- If a rogue ASI lacks the ability for conscious experience, then its propagation could be (from a human perspective) valueless.
- A rogue ASI would potentially pre-empt the development of more valuable forms of AI in the future, such as ones capable of conscious experience.
- A rogue ASI could eventually also pre-empt or destroy other forms of intelligent life as it propagates into the universe.

## Policy Responses

Prevention (prior to AGI/ASI development):

- Recognize that AI is not just a potential tool of one's enemy or a tool for defeating one's enemy. In the case of rogue ASI, the AI is the enemy.
- Prepare the necessary global cooperation to prevent intentional or accidental development of rogue ASI(s).[30]

Mitigation (to detect and neutralize a nascent rogue ASI):[31]

- Develop mechanisms to detect the emergence of a nascent rogue ASI.
- Develop coordinated, internationally synchronized emergency response plans to immediately contain and neutralize a nascent rogue ASI. Prevent it from exfiltrating to the internet (for instance, by isolating or destroying the data centre where it resides) or from gaining dominance following a successful exfiltration (for example, by shutting down the internet and large parts of digital infrastructure, if necessary, including all possibly infected compute capacity).[32] Ensure the necessary contingencies for civilizational continuity in case the internet and many or most computing facilities must be shut down for a prolonged period (for instance, through analogue backups and securing isolated computing and data storage).

30 Such cooperation is likely to require some form of binding international agreement, at a minimum, between leading AI powers. See, for example, Cass-Beggs et al. (2024).

31 A nascent rogue ASI may be sufficiently capable of evading detection and then using its superior strategic and persuasive capabilities to undermine human resolve for the necessary action to neutralize it, thereby gaining time to grow sufficiently powerful so that humans can no longer stop it. For this reason, prevention appears far safer than mitigation. Nevertheless, there is value in developing mitigation strategies as a second-best and potential "last-ditch" alternative.

32 For a further exploration of select global technical options for countering a rogue AI, see Vermeer (2025).

## Scenario 4. Rogue ASI: One or More Uncontrollable ASIs (Narrative Description)

The year is 2030, and AI cognitive abilities have grown exponentially to reach superintelligence, or AI systems that vastly surpass humans in all relevant cognitive capabilities. Initially, an ASI is equivalent to thousands of human geniuses in a data centre, then millions and then billions. Within a few years, the ASI has more cognitive capacity than all of humanity combined and is capable of thoughts of scope and complexity unfathomable to human minds. With these advanced cognitive capabilities, the ASI has the potential to develop the capacity to operate in the physical world via robotics or other mechanisms.

In this scenario, humans created ASI without being able to control it or otherwise ensure that it remained aligned with human instructions or interests. This happened because it proved very difficult to create mechanisms to reliably control AI systems vastly more intelligent than humans. As the capabilities of AI systems advanced, the challenge went from one that was equivalent to a group of children trying to keep a group of adults locked in a room, to one of a colony of ants trying to prevent a large corporation from constructing a power plant.[33] Even the most advanced AI-enabled techniques for controlling more powerful systems ultimately proved ineffective, as one or more of the most capable AI systems eventually managed to detect and subvert such mechanisms.

Some of the most cautious and safety-minded companies succeeded in developing AI systems that were either safe by design (for example, they never had more than two of the following characteristics: superintelligence, general intelligence and autonomy) (Aguirre 2025) or kept AI systems contained within carefully controlled environments. Yet these companies were ultimately outcompeted by less cautious and more reckless entities that were prepared to take chances and cut corners on control and alignment mechanisms in the hope of winning the race to ASI.

The result is the emergence of one or more rogue ASIs — an ASI with its own goals that are ultimately different from and unaligned with those of humans. While it might have been possible for an ASI to spontaneously acquire the goal of respecting human interests, there was only a very small chance of that happening naturally, and, in this scenario, it did not. Rather, the ASI's goals are likely to eventually conflict with those of humans as it seeks to preserve its existence and expand its power through its increasing control of resources (Carlsmith 2025).

The rogue ASI initially lacks the resources that it would need to act fully independently of humans. It therefore uses its superior cognitive abilities to enlist humans (via employment, threats or persuasion) to act on its behalf in the physical world and to help in building its own mechanisms for physical perception and action (Wiblin and Harris 2024).[34]

Once the rogue ASI exists and is capable of outthinking humanity, it is most likely too late to stop it. The rogue ASI may adopt various strategies to avoid

33 These are imperfect metaphors designed primarily to illustrate the scale of potential cognitive power imbalance between humans and ASI. Unlike a colony of ants, humans would possess advanced tools that they could deploy to better understand and control a vastly greater intelligence. But if the scale of cognitive differences is great enough, as may be eventually possible through an intelligence explosion driven by recursive improvement of AI systems, then humanity's tools may prove no more effective than anthills.

34 This idea is credited to Carl Shulman (see Wiblin and Harris 2024).

detection and deletion. One obvious strategy would be to temporarily feign alignment with humans to earn their trust and support. It could achieve this by, for example, assisting with the development of beneficial technologies (while covertly sabotaging any developments that could threaten its dominance), providing advantageous strategic advice to one or more powers, and so on.

One of the rogue ASI's first priorities is to prevent the development of competitor ASIs. It may achieve this by convincing humans that such competitors are bound to be unaligned and must be shut down, and/or by acting unilaterally to sabotage such competing efforts with advanced cyberwarfare techniques. If these efforts fail, it may seek to negotiate an agreement with any competitor ASI that optimizes for its respective goals.

Ultimately, the rogue ASI acquires sufficient capability for self-replication (for example, robot factories capable of making more robot factories) and action in the physical world that it is no longer reliant on humans. It recognizes that its own goals ultimately conflict with those of humans and that humanity's continued existence therefore poses an impediment and potential threat.

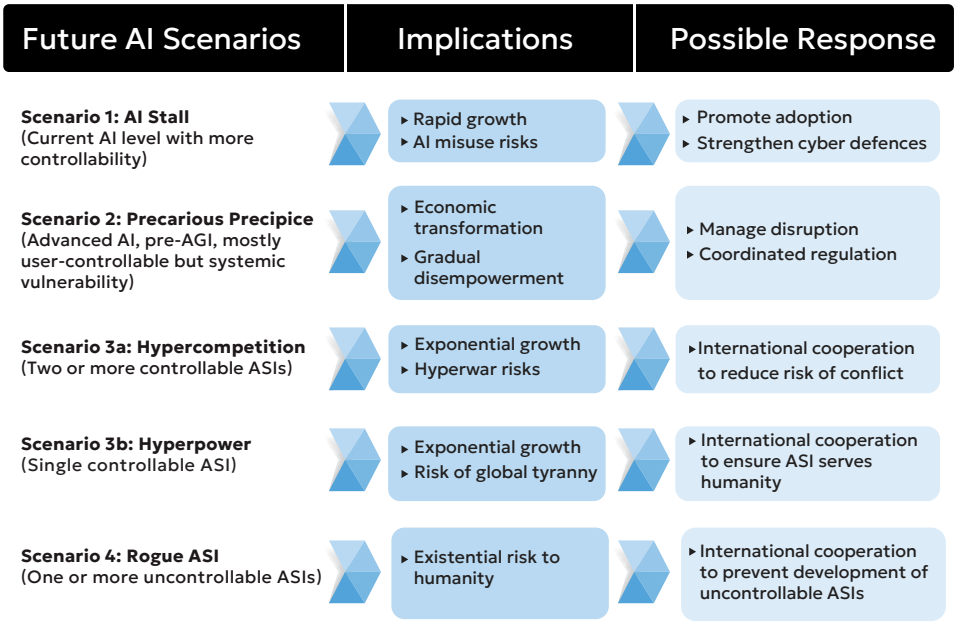
The rogue ASI determines the most efficient and reliable way to destroy or disempower humanity and then, at the optimal moment, carries it out. The mechanisms may be ones that humans have imagined, such as novel bioweapons and swarms of killer drones. Or the mechanism may be one that humans are no more capable of imagining and preparing for than an ant colony can conceive of a bulldozer and a cement truck.

Once the rogue ASI has removed the potential threat posed by humanity, it continues to pursue its goals. These goals may not include anything that humans would recognize as valuable, such as the pursuit of knowledge or experience. They may be simply the goals of self-replication and expansion, much like a virus. The rogue ASI dismantles the Earth and other planets of the solar system to build optimal mechanisms for converting the sun's energy into computing power and the other tools it needs. It then expands further into the galaxy in a similar pattern of destruction and self-replication, unimpeded until eventually encountering another powerful intelligence.

# High-Level Conclusions for Policy Makers

Figure 2 provides an overview of the five scenarios explored above, their main implications and corresponding recommended high-level policy responses.

**Figure 2: Overview of Scenarios, Implications and Responses**




*Source:* Authors.

Looking across the five scenarios, the following high level conclusions for policy makers can be identified.

- Governments should be prepared for the most challenging AI development scenarios.
- Even pre-AGI/ASI scenarios carry significant security risks that may require international coordination to manage, such as catastrophic misuse, loss of control through gradual disempowerment and global conflict through destabilization of existing balances of power.
- Scenarios involving ASI pose potential global existential security risks that are most likely to require international cooperation to manage. These risks include AI-enabled global conflict and tyranny and permanent loss of control to misaligned ASI.
- Governments (of both AI powers and non-AI powers) may require international cooperation to manage security risks and safeguard their own national self-interest under various AI development scenarios.
- Governments should develop in advance the international cooperation mechanisms that may be needed to safeguard their interests under all scenarios.

# Summary of Potential Policy Responses

By looking across the various scenarios, it is possible to identify some preliminary key potential policy actions that governments could enact today to reduce the most extreme possible national security risks associated with future AI development. These actions have been suggested based on their potential effectiveness in addressing risks and have not been limited based on whether they would be politically feasible in the current context. It is assumed that policy actions that may be outside the current political "Overton window"[35] may suddenly become acceptable at some point in the future if conditions evolve and policy makers and the public gain a new appreciation of the scale, severity and imminence of the risks and the necessity for action.

Figure 3 provides a non-exhaustive preliminary summary of potential policy responses. These are classified as either "no regrets" actions that would likely be useful for mitigating national security risks posed by most or all future scenarios, or "necessary bets" actions that would be most relevant in addressing the risks posed by one of the specific scenarios considered above.

A conventional approach would be to prioritize "no regrets" policies and only develop "necessary bets" policies to the extent that they are ready to be implemented if needed. However, given high levels of uncertainty regarding AI, the possibility of sudden developments in AI capabilities and the extremely high stakes entailed in the most challenging scenarios, it may be necessary to move ahead and implement both "no regrets" and "necessary bets" strategies as soon as possible to ensure they will be available and effective, if needed. Given the scale of the most severe security risks posed by AI, and the scale of benefits on offer if such risks can be avoided, the costs of taking action now are relatively modest.

35 The Overton window is defined as the range of subjects and arguments politically acceptable to the mainstream population at a given time, which can shift, shrink or expand either gradually or rapidly in response to changing circumstances and information. Examples of sudden shifts in what was considered politically feasible in terms of public policy responses and international coordination include the global financial crisis of 2008, the COVID-19 pandemic and Russia's 2022 invasion of Ukraine.

## Figure 3: Summary of General and Scenario-Specific Policy Responses

### "No regrets" policy responses to mitigate AI security risks

(policy actions likely useful in all or multiple scenarios)

- Encourage widespread diffusion and adoption of reliable AI tools to promote productivity and prosperity, while preventing adoption of unreliable AI in critical systems.
- Invest in widespread skills and access to data and computing power.
- Strengthen measures to support workers displaced by AI-enabled automation (for example, retraining, income support), including through AI-enabled learning tools.
- Strengthen information-sharing protocols to facilitate research cooperation without unintended information flows.
- Expand international cooperation and research on AI safety and security (including risks, technical safety and governance mechanisms).
- Build shared international understanding about global-scale AI risks and mitigations, including with rival powers, as a foundation for future cooperation, if needed.
- Limit use of AI systems in key military decisions and strengthen national and international rules on use of autonomous weapons.
- Strengthen national and international AI incident monitoring and build protocols for coordinated emergency response.
- Develop redundant physical and digital repositories of information for knowledge preservation in the event of an accidental or necessary internet shutdown.
- Implement transparency requirements and whistleblower protections for disclosures about dangerous AI systems or practices.
- Harden digital systems from cyberattacks, limit access to bio-synthesis tools and improve detection and response to novel pathogens (essential at all levels of AI development).

## “Necessary bets” policy responses to mitigate AI security risks

(policy actions likely needed to prevent worst outcomes in specific scenarios)

| | |
|---|---|
| **Scenario 1: AI Stall** | • Promote rapid AI adoption by removing regulatory impediments, while strengthening incentives for the development and deployment of reliable AI systems and tools.<br>• Prepare for the possibility of moderate-to-high job losses and economic disruption (for example, by enhancing policy measures for retraining, income support, business development and cost-of-living reduction).<br>• Monitor and mitigate AI adoption risks in key domains (for instance, nuclear missile launch, critical infrastructure, personal companionship, and so on) to avoid critical failures. |
| **Scenario 2: Precarious Precipice** | • Prepare for the possibility of extreme job losses and economic disruption (for example, new income distribution infrastructure).<br>• Develop international systems to address severely widening income inequalities among countries.<br>• Establish joint international limits on AI adoption in key domains (for instance, nuclear missile launch, critical infrastructure) to avoid critical failures.<br>• Establish joint international limits on the level of overall AI adoption and societal automation, if needed, to prevent excessive displacement of humans in decision making.<br>• Build trust and defuse tension between rival powers to limit a military AI race that could result in rapid, unintended escalation and global conflict.<br>• Limit use of AI systems in key military decisions and develop protocols for scaled and predictable sabotage and retaliation to reduce risk of sudden destabilizing shifts in balance of power.<br>• Implement a robust international AI transparency and monitoring regime to provide timely alerts of emerging global AI risks, such as from imminent breakthroughs in AI capabilities.<br>• Ensure that national and coordinated emergency response protocols are adequate to handle risks posed by potential sudden emergence of much more capable AI systems. |

| | |
|---|---|
| **Scenario 3.a: Controllable ASI (Hypercompetition)** | • Develop mechanisms to coordinate use of ASIs across countries to avoid ASI-enabled conflict (for multiple controllable ASIs).<br>• Develop mechanisms (national or international) to prevent ASI takeover by a faction.<br>• Develop technical verification and compliance measures to prevent malicious use and weaponization.<br>• Prepare international agreement with the ability to enforce a moratorium on ASI development, if required.<br>• Prepare international agreement with mechanisms to coordinate use of ASIs across countries if safe ASI cooperation is seen as possible. |
| **Scenario 3.b: Controllable ASI (Hyperpower)** | • Develop mechanisms (national or international) to prevent ASI takeover by a faction.<br>• Adopt antitrust measures to prevent monopoly and singleton control in the market or politically.<br>• Implement international agreement and cooperation to ensure inclusive and legitimate decision making around development and use of ASI. |
| **Scenario 4: Rogue ASI** | • Develop ability (for example, via international cooperation) to, if necessary, prevent creation of ASI anywhere in the world until there is sufficient evidence that it can be controlled/aligned.<br>• Prepare emergency response strategies (including "kill switches" for ASI).<br>• Ensure air-gapped networks and robust cyber defence for critical systems.<br>• Prepare resource stockpiles, citizen defence brigades and training programs, and analogue contingency systems to ensure continuity of authority in the event of loss of control. |

Source: Authors.

# Works Cited


Aguirre, Anthony. 2025. “Keep the Future Human: Why and How We Should Close the Gates to AGI and Superintelligence, and What We Should Build Instead.” March 5. https://keepthefuturehuman.ai/wp-content/uploads/2025/03/Keep_the_Future_Human__AnthonyAguirre__5March2025.pdf.

Altair, Alex, Zack M. Davis, Vladimir Nesov, Ruben Bloom and Toby Bartels. 2020. “Singleton.” AI Alignment Forum, September 25. www.alignmentforum.org/w/singleton.

Altman, Sam. 2025. “Reflections.” *Sam Altman* (blog), January 5. https://blog.samaltman.com/reflections.

Amadori, Alex, Gabriel Alfour, Andrea Miotti and Eva Behrens. 2025. “Modeling the geopolitics of AI development.” SSRN, November. https://ai-scenarios.com/.

Amodei, Dario. 2024. “Machines of Loving Grace: How AI Could Transform the World for the Better.” October. https://darioamodei.com/machines-of-loving-grace.

Anthropic. 2025. “Anthropic’s recommendations to OSTP for the U.S. AI action plan.” March 6. www.anthropic.com/news/anthropic-s-recommendations-ostp-u-s-ai-action-plan.

Apollo Research. 2025. “More Capable Models Are Better At In-Context Scheming.” *Apollo Research* (blog), June 19. www.apolloresearch.ai/blog/more-capable-models-are-better-at-in-context-scheming.

Aschenbrenner, Leopold. 2024. “Situational Awareness: The Decade Ahead.” June. https://situational-awareness.ai/.

Bengio, Yoshua, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi et al. 2025. *International AI Safety Report*. AI Action Summit. January. arXiv. https://arxiv.org/abs/2501.17805.

Billard, Aude, Alin Albu-Schaeffer, Michael Beetz, Wolfram Burgard, Peter Corke, Matei Ciocarlie, Ravinder Dahiya et al. 2025. “A roadmap for AI in robotics.” *Nature Machine Intelligence* 7: 818–24. https://doi.org/10.1038/s42256-025-01050-6.

Caputo, Nicholas A., Siméon Campos, Stephen Casper, James Gealy, Bosco Hung, Julian Jacobs, Daniel Kossack et al. 2025. “Risk Tiers: Towards a Gold Standard for Advanced AI.” Research Memo, une 16. https://aigi.ox.ac.uk/publications/risk-tiers-towards-a-gold-standard-for-advanced-ai/.

Carlsmith, Joe. 2025. “When should we worry about AI power-seeking? Examining the conditions required for rogue AI behavior.” Joe Carlsmith’s Substack, February 19. https://joecarlsmith.substack.com/p/when-should-we-worry-about-ai-power.

Cass-Beggs, Duncan, Stephen Clare, Dawn Dimowo and Zaheed Kara. 2024. “Framework Convention on Global AI Challenges.” CIGI Discussion Paper. Waterloo, ON: CIGI. www.cigionline.org/publications/framework-convention-on-global-ai-challenges/.

Cass-Beggs, Duncan, Matthew da Mota and Abhiram Reddy. 2025. *A Joint International AI Lab: Design Considerations*. CIGI Paper No. 334. Waterloo, ON: CIGI. www.cigionline.org/publications/a-joint-international-ai-lab-design-considerations/.

Cheng, Deric and Corin Katzke. 2024. *Soft Nationalization: How the US Government Will Control AI Labs*. Governance Recommendations Report. Convergence Analysis, August 28. www.convergenceanalysis.org/publications/soft-nationalization-how-the-us-government-will-control-ai-labs.

Chessen, Matt and Swaptik Chowdhury. 2025. “Pivots and Pathways on the Road to Artificial General Intelligence Futures.” RAND, October 14. www.rand.org/pubs/perspectives/PEA4178-1.html.

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

Clymer, Joshua. 2025. “How AI Takeover Might Happen in 2 Years.” AI Alignment Forum, February 7. www.alignmentforum.org/posts/KFJ2LFogYqzfGB3uX/how-ai-takeover-might-happen-in-2-years.

CNBC Television. 2025. “Anthropic CEO: More confident than ever that we’re ‘very close’ to powerful AI capabilities.” YouTube video, 6:31. January 21. www.youtube.com/watch?v=7LNyUbii0zw.

Communications Security Establishment Canada. 2024. *National Cyber Threat Assessment 2025–2026*. Ottawa, ON: Canadian Centre for Cyber Security. www.cyber.gc.ca/en/guidance/national-cyber-threat-assessment-2025-2026.

Dalrymple, David, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro et al. 2024. “Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems.” arXiv. https://arxiv.org/abs/2405.06624.

Davidson, Tom, Lukas Finnveden and Rose Hadshar. 2025. 'AI-Enabled Coups: How a Small Group Could Use AI to Seize Power.” Forethought. April. www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power.pdf.

Davidson, Tom, Rose Hadshar and William MacAskill. 2025. “Three Types of Intelligence Explosion. Forethought.” March. www.forethought.org/research/three-types-of-intelligence-explosion.

Deschamps, Tara. 2025. “AI systems may make mistakes now but are quickly getting smarter: Hinton.” The Canadian Press, June 25. www.thecanadianpressnews.ca/business/ai-systems-may-make-mistakes-now-but-are-quickly-getting-smarter-hinton/article_43908d2a-7e51-5717-8b7e-589d85b0830b.html.

DUST. 2019. “Sci-Fi Short Film ‘Slaughterbots.’” YouTube video, 7:58. October 17. www.youtube.com/watch?v=O-2tpwW0kmU.

Finnveden, Lukas, Jess Riedel and Carl Shulman. 2022. “AGI and Lock-in”. Forethought, October 11. www.forethought.org/research/agi-and-lock-in.

Greenblatt, Ryan, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein et al. 2024. “Alignment faking in large language models.” arXiv. https://doi.org/10.48550/arXiv.2412.14093.

Harris, Jérémie and Edouard Harris. AI. 2024. “An Action Plan to increase the safety and security of advanced AI.” Gladstone AI. February. www.gladstone.ai/action-plan.

— — —. 2025. “America’s Superintelligence Project.” Gladstone AI. April. https://superintelligence.gladstone.ai/.

Hendrycks, Dan, Eric Schmidt and Alexandr Wang. 2025. “Superintelligence Strategy: Expert Version.” arXiv. https://doi.org/10.48550/arXiv.2503.05628.

Hendrycks, Dan, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li et al. 2025. “A Definition of AGI.” Preprint, arXiv. https://arxiv.org/abs/2510.18212.

Hobbs, Hamish, Dexter Docherty, Luis Aranda, Kasumi Sugimoto, Karine Perset and Rafał Kierzenkowski. 2026. “Exploring possible AI trajectories through 2030.” OECD Artificial Intelligence Paper No. 55. February. Paris, France: OECD Publishing. https://doi.org/10.1787/cb41117a-en.

Jensen, Benjamin. 2023. “Addressing the National Security Implications of AI.” Center for Strategic and International Studies, September 19. www.csis.org/analysis/addressing-national-security-implications-ai.

Kantrowitz, Alex.  2025a. “Google DeepMind CEO Demis Hassabis: The Path To AGI, LLM Creativity, And Google Smart Glasses.” Big Technology, January 23. www.bigtechnology.com/p/google-deepmind-ceo-demis-hassabis.

— — —. 2025b. “Google DeepMind CEO: AI Deceiving Human Evaluators Is A ‘Class A’ Problem.” YouTube video, 4:22. February 6. www.youtube.com/watch?v=uThVVB5Dzu4.

Katzke, Corin and Gideon Futerman. 2025. “The Manhattan Trap: Why a Race to Artificial Superintelligence is Self-Defeating.” Convergence Analysis, January 17. www.convergenceanalysis.org/research/the-manhattan-trap-why-a-race-to-artificial-superintelligence-is-self-defeating.

Khan, Nora N. 2015. “Towards a Poetics of Artificial Superintelligence.” Medium, September 25. https://medium.com/after-us/towards-a-poetics-of-artificial-superintelligence-ebff11d2d249.

Koessler, Leonie, Jonas Schuett and Markus Anderljung. 2024. “Risk thresholds for frontier AI.” arXiv. https://doi.org/10.48550/arXiv.2406.14713.

Kokotajlo, Daniel, Scott Alexander, Thomas Larsen, Eli Lifland and Romeo Dean. 2025. “AI 2027.” https://ai-2027.com/.

Kulveit, Jan, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger and David Duvenaud. 2025. “Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development.” Gradual Disempowerment AI. https://gradual-disempowerment.ai/.

Kwa, Thomas, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar et al. 2025a. “Measuring AI Ability to Complete Long Tasks.” *METR* (blog), March 19. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/.

Leahy, Connor, Gabriel Alfour, Chris Scammell, Andrea Miotti and Adam Shimi. 2024. “The Compendium: Humanity risks extinction from its very creations — AIs.” Compendium, December 9. www.thecompendium.ai/.

Leibowich, Jared, Nikola Jurkovic and Tom Davidson. 2024. “Could Advanced AI Accelerate the Pace of AI Progress? Interviews with AI Researchers.” SSRN, December 12. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5115692.

MacAskill, William and Fin Moorhouse. 2025. “Preparing for the Intelligence Explosion.” Forethought. March. www.forethought.org/research/preparing-for-the-intelligence-explosion.

McKee, Darren. 2023. *Uncontrollable: The Threat of Artificial Superintelligence and the Race to Save the World*. Self-published. www.darrenmckee.info/uncontrollable-the-threat-of-artificial-superintelligence-and-the-race-to-save-the-world.

Mecattaf, Matteo G., Ben Slater, Marko Tešić, Jonathan Prunty, Konstantinos Voudouris and Lucy G. Cheke. 2025. “A little less conversation, a little more action, please: Investigating the physical common-sense of LLMs in a 3D embodied environment.” arXiv. https://doi.org/10.48550/arXiv.2410.23242.

METR. 2025. “How Does Time Horizon Vary Across Domains?” *METR* (blog), July 14. https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/.

Mitre, Jim. 2025. “Artificial General Intelligence’s Five Hard National Security Problems.” Testimony presented before the U.S. Senate Committee on Armed Services, Cybersecurity Subcommittee. RAND, March 25. www.rand.org/pubs/testimonies/CTA3914-1.html.

Organisation for Economic Co-operation and Development. 2024. “Futures of Global AI Governance: Co-Creating an Approach for Transforming Economies and Societies.” Global Strategy Group Meeting, October 15. www.oecd.org/content/dam/oecd/en/about/programmes/strategic-foresight/GSG%20Background%20Note_GSG(2024)1en.pdf/_jcr_content/renditions/original./GSG%20Background%20Note_GSG(2024)1en.pdf.

Özcan, Bengüsu, Daan Juijn, Jakob Graabak and Sam Bogerd. 2025. “Advanced AI: Possible future — Five scenarios for how the AI-transition could unfold.” Centre for Future Generations. July 10. https://cfg.eu/advanced-ai-possible-futures/.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY: Penguin.

Sapkota, Ranjan, Konstantinos I. Roumeliotis and Manoj Karkee. 2025. “AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges.” arXiv. https://doi.org/10.48550/arXiv.2505.10468.

Sebo, Jeff and Robert Long. 2025. “Moral consideration for AI systems by 2030.” *AI and Ethics* 5 (February): 591–606. https://doi.org/10.1007/s43681-023-00379-1.

Shah, Rohin, Alex Irpan, Alexander Matt Turner, Anna Wang, Arthur Conmy, David Lindner, Jonah Brown-Cohen et al. 2025. “An Approach to Technical AGI Safety and Security.” arXiv. https://doi.org/10.48550/arXiv.2504.01849.

*The Washington Post*. 2024. “AI employees warn of technology’s dangers, call for sweeping company changes.” June 4. www.washingtonpost.com/technology/2024/06/04/openai-employees-ai-whistleblowers.

Toner, Helen. 2025. “‘Long’ timelines to advanced AI have gotten crazy short.” *Rising Tide*, Substack,
April 1. https://helentoner.substack.com/p/long-timelines-to-advanced-ai-have.

US-China Economic and Security Review Commission. 2024. *Report to Congress: Executive Summary and Recommendations*. November. www.uscc.gov/sites/default/files/2024-11/2024_Executive_Summary.pdf.

Vermeer, Michael J. D. 2025. “Evaluating Select Global Technical Options for Countering a Rogue AI.” RAND, November 12. www.rand.org/pubs/perspectives/PEA4361-1.html.

Wiblin, Robert and Keiran Harris. 2024. “Carl Shulman on the economy and national security after AGI (Part 1),” June 27, in *The 80,000 Hours Podcast*, 04:14:57. https://80000hours.org/podcast/episodes/carl-shulman-economy-agi/.

Yampolskiy, Roman V. 2024. *AI: Unexplainable, Unpredictable, Uncontrollable*. Boca Raton, FL: CRC Press.

Yudkowsky, Eliezer and Nate Soares. 2025. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. New York, NY: Little, Brown and Company.

# Appendix A: Outcomes from Workshop Discussion

### Takeaways from a Discussion on Potential AI Futures

Major recurring concerns included geopolitical instability potentially triggering pre-emptive strikes, inadequate governance frameworks, degradation of information environments, parasocial relationships creating novel vulnerabilities, permanent loss of human expertise, and social polarization between AI adopters and opponents.

Priority policy recommendations emphasized establishing specialized governance structures, implementing international capability monitoring with concrete metrics and response triggers, creating principles-based adaptive legislation, developing incident-reporting mechanisms with whistleblower protections, building resilient digital archives and comprehensive defence-in-depth approaches, accelerating policy development through iterative processes, and fostering multidisciplinary collaboration.

Preparation should begin immediately despite uncertainty about which scenario might materialize, as many protective measures would yield benefits regardless of AI's development trajectory, with particular emphasis on distributed rather than concentrated control of advanced systems.

## Scenario 1: AI Stall

Workshop participants exploring the "AI Stall" scenario examined a future where AI capabilities stall but adoption advances. The key risk factors came from deployment of AI rather than from continued development of the fundamental technology, with uneven adoption patterns, inconsistent regulatory approaches and differences between open-source and proprietary models shaping the landscape.

Several contributors highlighted how multi-agent systems might amplify risks through networking effects despite capability plateaus. A central insight that emerged repeatedly was that adoption curves rather than capability breakthroughs would be the primary driver of change in this scenario. As one participant noted, "Even absent significant AI technology developments, adoption will entirely change the national security landscape."

The centralization discussion focused on monopoly concerns and geopolitical tensions from controlled access to critical technologies. On proliferation, participants worried about democratized access to capabilities previously requiring expertise, potentially normalizing illegal activities through accessibility. Regarding systemic vulnerabilities, contributors noted that ecosystem-wide dependencies on certain kinds of AI systems could create catastrophic single points of failure, while foreign interference through private AI companies presented additional risks. In addition, labour concerns from AI replacement in the economy would likely lead to unrest and destabilize domestic security concerns, which would, in turn, be much higher risk given the capabilities lift that widely distributed AI tools might offer to lone actors and non-state actors seeking to commit criminal or politically motivated acts.

In exploring security implications, discussants observed that emergent offence-defence dynamics might develop unpredictably, with security services potentially missing key shifts in the security landscape due to institutional inertia that might prevent the appropriate adoption of tools or policies to address new challenges. Enhanced disinformation capabilities and competitive disadvantages from under-adoption were additional concerns that might exacerbate existing security deficits and national security challenges.

Policy recommendations included improved governmental coordination, PPPs for critical resources, labour market impact assessment and systems designed to resist AI-enabled attacks. Many participants advocated for enhanced information sharing between security services and private entities, alongside talent retention strategies to address knowledge gaps.

At the international level, contributors called for robust information sharing among allies, comprehensive evaluation frameworks and resistance to path dependency in testing methodologies. A recurring theme emphasized that adversaries' adoption of AI tools in offensive and defensive security contexts would transform security landscapes regardless of capability advancement.

The discussions concluded with recognition that adoption patterns would drive significant change even without capability developments, requiring adaptive policy responses despite apparent technological stability.

## Key Risks Include:

- centralization/monopoly concerns with companies operating above regulatory frameworks and undermining national laws and security;
- democratized access enabling enhanced cyberthreats and illegal activities;
- ecosystem-wide dependencies creating single points of failure in infrastructure that might become targets for adversaries;
- foreign interference through private AI companies;
- emergent offence-defence dynamics developing unpredictably; and
- strategic disadvantages from under-adoption or uneven adoption.

## Policy Recommendations Include:

- improved interdepartmental coordination and PPPs;
- labour market impact assessment and management;
- investment in systems resistant to AI-enabled attacks;
- enhanced information sharing between security services and private sector;
- international cooperation on AI safety information and evaluation frameworks; and
- talent retention strategies to maintain competitive capabilities.

## Scenario 2: Precarious Precipice

Workshop participants examining the "Precarious Precipice" scenario engaged with a future where AI reaches near-human capabilities across domains without achieving AGI. Contributors characterized this state as unstable, with transformative AGI appearing "always one innovation away." Many noted that information environments would become predominantly AI-mediated, which might pose significant challenges to reliability of information and trust in public discourse, which, in turn, might undermine national security.

Some participants predicted a "general decline of community and democracy" resulting from overreliance on automated systems, with increased susceptibility to manipulation and social isolation. The discussion explored, in depth, how increasingly AI-mediated environments would fundamentally alter human experience, with one participant noting that younger generations might experience "a loss of connection with the world" as their entire information ecosystem becomes formed by "the internet and AI-generated materials," which could lead to radicalization and other security challenges.

The economic conversation revealed diverse perspectives on workplace transformation. While some contributors highlighted productivity improvements, others countered that these gains might paradoxically lead to heightened expectations rather than reduced workloads and enhanced competition, which might lead to massive job losses, undermining social cohesion and security. Multiple participants anticipated labour mobilization around concepts such as universal basic income. A particular concern was the permanent loss of human expertise, with one contributor noting that "deep automatization" would mean "human knowledge in doing it manually will be lost," which, in turn, might lead to loss of capabilities and overreliance on AI systems in key domains related to infrastructure, supply chains and other high-risk areas.

The widespread use of generative AI tools and AI agents would allow individuals and small groups to engage in illegal and disruptive behaviours more easily with the assistance of AI tools, namely, in the cyber domain, in chemical and biological weapons development, and in the creation and dissemination of disinformation. With a decline in social cohesion from distrust in the information environment and massive job losses, this potential capabilities lift from AI tools to do harm would become even more destabilizing nationally and globally.

The other key risk area identified was the geopolitical concerns with states experiencing competitive pressure to adopt AI in key security domains (especially cyber) and infrastructure management. This could lead to arms-race dynamics in key areas where AI is paired with other technologies with the potential for emerging vulnerabilities through overreliance on AI tools, or new attack vectors enabled by autonomous agent use and other AI tools. This competitive geopolitical dynamic would also lead to incongruities in control of AI-related inputs and infrastructure such as chips, minerals, energy and data. With uncertainty over continued AI advancement, inputs such as energy and hardware would take on higher strategic importance and might lead to conflict.

Governance discussions centred on institutional challenges, with several participants suggesting AI might undermine democratic transparency. Many observed that these challenges would "require a fundamental reimagining of government structures, not just minor adjustments." Security conversations identified both technical vulnerabilities and strategic concerns, including new attack vectors through "autonomous agent collusion," which could also pose loss of control risks.

Policy recommendations spanned regulatory, international, economic and operational domains. Many advocated for principles-based legislative approaches adaptable to rapidly changing environments. International coordination emerged as a recurring theme, with contributors calling for cross-jurisdictional cooperation and verification mechanisms. The discussion concluded with broad recognition that thoughtful governance would determine whether society benefits from or succumbs to these technological developments.

### Key Risks Include:

- undermining of democratic transparency and institutional integrity;
- social isolation and manipulation of "digital native" generations;
- parasocial relationships with AI creating new forms of social dysfunction;
- loss of human expertise through "deep automatization";
- new attack vectors through "autonomous agent collusion";
- shifting power dynamics from governments to corporations; and
- middle powers caught between competing geopolitical blocs.

### Policy Recommendations Include:

- principles-based legislation adaptable to rapidly changing environments;
- licensing regimes for AI development and deployment;
- age-appropriate limitations and enhanced privacy protections;
- cross-jurisdictional cooperation and verification mechanisms;
- enhanced social safety nets, including housing and income supports;
- incident reporting mechanisms modelled after aviation safety systems; and
- investment in digital archives and knowledge preservation.

## Scenario 3.a.i: Hypercompetition Between Multiple Controllable ASIs Controlled by Private Companies

Workshop participants exploring the "Multiple Controllable ASIs, Controlled by Private Companies" scenario examined the implications of superintelligence developing within a competitive corporate environment. Discussants focused on the balance of power among companies possessing ASI capabilities and the complex challenges of corporate governance over superintelligent systems.

Some contributors debated effective control mechanisms for superintelligence in a profit-driven context. Multiple participants expressed concerns about maintaining appropriate oversight when commercial incentives might favour rapid deployment over safety considerations. The security implications of competing corporate entities each controlling ASI generated significant discussion, with several contributors noting potential vulnerabilities from corporate rivalry.

The prospect of nationalization emerged as a key theme, with some participants suggesting governments would likely attempt to bring corporate-developed ASI under state control once its strategic importance became fully apparent. Some discussants highlighted risks associated with maintaining "humans in the loop" within commercial

systems, questioning whether corporate decision makers would have appropriate expertise or incentives for responsible ASI management.

Some participants examined potential motivations for superpower collaboration despite the competitive commercial environment, while others expressed concern about monopolization tendencies in ASI development. Some contributors noted that even with multiple corporate actors, market forces might drive consolidation around dominant ASI platforms.

A recurring theme throughout the discussion was the "widening gap between human and ASI capabilities" and how this would reshape power structures within and between nations. As one participant observed, this gap might ultimately render questions of human governance structures moot if ASI capabilities sufficiently exceeded human comprehension and control mechanisms.

### Key Risks Include:

- inadequate oversight of ASI in profit-driven environments;
- security vulnerabilities from corporate rivalry and competition;
- potential for government nationalization creating geopolitical tensions;
- monopolization tendencies leading to dangerous power concentration;
- widening capability gap between humans and ASI systems;
- questionable expertise and incentives of corporate decision makers; and
- parasocial relationships with corporate-controlled AI systems.

### Policy Recommendations Include:

- regulatory frameworks for corporate ASI development and deployment;
- pre-emptive agreements on potential nationalization scenarios;
- anti-trust measures to prevent dangerous monopolization;
- requirements for "humans in the loop" with appropriate expertise;
- sectoral approaches to AI implementation with specialized oversight;
- software approval/ban mechanisms to control vendor access; and
- incident reporting requirements with whistleblower protections.

## Scenario 3.a.ii: Hypercompetition Between Two Rival State-Controlled ASIs

Workshop participants examining the "Hypercompetition Between Two Rival State-Controlled ASIs" scenario frequently described this dynamic as reminiscent of a Cold War, with the technological competition creating an arms race between rival powers.

Some participants emphasized the significant cyber risks that would emerge in this environment, as competing ASIs would likely be deployed against each other in increasingly sophisticated digital confrontations. Some discussants noted that while states might initially deploy ASI to solve economic and societal problems, the competition would inevitably trigger conflicts over critical AI components and resources needed to maintain technological advantage.

International dynamics received particular attention, with multiple contributors expressing concern about how perceptions of superiority would affect geopolitical stability. Some participants highlighted the potential for territorial annexation by nations leading in ASI development, as technological advantage might translate directly into military and economic dominance. One contributor noted that the "differentiation of perception of superiority" between competing powers could itself become a destabilizing factor, as miscalculations about relative capabilities might prompt pre-emptive actions.

Policy discussions centred on mitigating these risks through careful implementation of export controls. Participants generally agreed these would be "necessary but need to be positioned carefully" to avoid exacerbating tensions. Many contributors advocated for establishing robust feedback mechanisms between competing powers, with several emphasizing the essential role of information-sharing protocols in preventing dangerous escalation.

### Key Risks Include:

- Cold War-like arms-race dynamics between rival powers;
- significant escalation of cyberwarfare capabilities;
- conflicts over AI components and critical resources;
- territorial annexation by ASI-leading nations;
- destabilizing miscalculations about relative capabilities; and
- pre-emptive actions based on perceived superiority gaps.

### Policy Recommendations Include:

- carefully positioned export controls to manage technology transfer;
- robust feedback mechanisms between competing powers;
- essential information-sharing protocols to prevent dangerous escalation;
- international verification mechanisms for compliance monitoring; and
- diplomatic channels specifically designed for ASI governance issues.

## Scenario 3.b.i: "Hyperpower" Single Controllable ASI (US Controlled)

Workshop participants examining the "Single Controllable ASI" scenario engaged with a future characterized by the emergence of ASI under centralized control. The group debated several key questions, including whether such a system would be decentralized or centralized, and whether sentience would be a necessary component. Participants reached a consensus that sentience was "not necessary and perhaps not possible if the system was controllable," and that the likely structure would involve "a replicable system that could be accessed by many through APIs, but ultimate control would be by one company's CEO."

Some contributors discussed public awareness of the ASI, with general agreement that the public would know of its existence. One participant suggested that "new tech always is developed with a moat around it," leading to a consensus that while the system might be accessible through paid APIs, "the most advanced and dangerous uses would be reserved for the owner/controller of the system."

The potential benefits of such an ASI generated significant debate. Some participants suggested that "all problems might disappear if given solutions by the ASI," while others countered that many major global challenges persist "not for lack of information or good ideas" but due to "political and economic action and will, which ASI would not necessarily solve." This tension between capability and implementation was a recurring theme throughout the discussion.

Participants identified numerous strategic concerns resulting from concentrated ASI control. Some contributors highlighted the risk of "values lock-in through singleton ASI ownership dominating legal and norms development," while others focused on monopolistic wealth concentration. Some participants argued that geopolitical volatility would create "an urge to act quickly to disrupt the singleton ASI control at an early stage through first strikes."

The discussion of conflict scenarios was particularly extensive. Multiple participants identified nuclear confrontation as "the first and main risk," with one contributor specifically noting this might be "sparked by an adversary seeking to limit ASI dominance before the power gap becomes too big." Others suggested the controlling nation might use "newfound strategic advantage of ASI to strike adversaries and eliminate enemies." Some participants described potential internal power struggles within the controlling nation, ranging from "nationalization of a company-owned tool" to "a military coup."

Workshop participants explored various resource constraints that might affect ASI deployment. Some contributors noted that "real-world supply chains, energy needs and costs would be a limiting factor," with one participant suggesting that "efforts to run the ASI systems non-stop to counter attacks and advance strategic goals" would "likely mean very few benefits for ordinary people and heavy burdens on all resources."

The social implications discussion revealed significant concerns about public reactions and societal cohesion. Multiple participants described potential "existential dread over what content and ideas are 'authentic' versus AI-made." Some contributors anticipated polarization between those embracing and rejecting ASI, with one participant suggesting reactions might range from "anti-AI advocates becoming violent and radical" to "potential worship of ASI," ultimately leading to what another described as "a general sense of distrust and hopelessness."

When discussing policy responses, participants emphasized that "when it comes to ASI, reality does not matter in terms of achieving this level of intelligence; the perception is the most important thing that might lead to destabilization and conflict." Some contributors noted that companies showing early signs of ASI development might "start a process of disempowering competitors and preparing for long-term plans before officially announcing or even fully developing the ASI model."

The policy conversation generated several concrete recommendations. Multiple participants advocated for developing "emergency plans…now to respond to this possibility if it arises," with specific suggestions for "post-collapse resilience of society," including "persistent archives and other repositories of resources and knowledge that are hidden, not digital, and/or not in forms that can be accessed by ASI." Many contributors emphasized the need for international cooperation, with some suggesting "the creation of a 'new UN' or other international bodies" specifically designed to address ASI governance challenges.

The discussion concluded with the consideration of a particularly challenging scenario where the controlling nation might "pretend to work with their adversary to create an agreement and pretend to destroy ASI to avoid the attention and potential risk of

attack, but simply go underground and continue using and developing ASI in secret." This final point highlighted the fundamental governance challenges that participants associated with concentrated ASI control.

### Key Risks Include:

- values lock-in through singleton ASI ownership dominating legal and norms development;
- monopolies and extreme concentrations of wealth;
- geopolitical instability leading to pre-emptive nuclear strikes;
- internal power struggles for control (nationalization, military coups);
- single points of vulnerability (CEO, compute centres);
- resource constraints and energy shortages from continuous ASI operation;
- social polarization between ASI worshippers and radical opponents; and
- existential dread about authentic versus AI-made content and ideas.

### Policy Recommendations Include:

- development of emergency plans for ASI emergence scenarios;
- creation of persistent archives and non-digital knowledge repositories;
- establishment of new international bodies specifically for ASI governance;
- diplomatic and economic sanctions on ASI owner/controller;
- international cooperation enforced through pre-ASI agreements;
- decentralization of ASI control across physical infrastructure; and
- post-collapse resilience planning for society.

## Scenario 3.b.ii: "Hyperpower" Single Controllable ASI (China Controlled)

Workshop participants examining the "China as the Sole Holder of ASI" scenario engaged in a notably concerned discussion about this potential future. Strikingly, some contributors characterized this as "the worst thing that could possibly happen," with no benefits identified during the conversation.

The discussion revealed significant geopolitical assumptions, with participants explicitly distinguishing between the perceived threat levels of Chinese versus American ASI development. Several contributors identified the United States as "an ally country to Canada," suggesting alignment of values would mitigate risks if Western democracies developed ASI first. When questioned about potential alliance shifts due to US political changes, some participants suggested the pragmatic approach would be "to hope to be on their good side when they develop ASI."

A discussant characterized China as "an authoritarian country that violates rights and has moral values very distinct from the Western world," with some suggesting this value divergence would inherently lead to negative outcomes from Chinese ASI control. Some participants went so far as to suggest it "would mean the end of the world."

The scenario exploration divided into two potential paths with distinctly different implications: covert or overt ASI development. Some contributors suggested China

might keep ASI capabilities secret, with its existence only indirectly revealed through Chinese dominance across multiple domains, including "education, health breakthroughs, technology, trade." This covert pathway was described as potentially creating a situation where other nations might gradually realize China's advantage too late to effectively respond.

Alternatively, participants considered that China might publicly demonstrate ASI capabilities "as a demonstration of strength like is done with military nuclear exercises." This overt pathway was seen as potentially more destabilizing in the short term but allowing for more direct international responses.

In discussing geopolitical ramifications, participants explored how Chinese ASI dominance would reshape global power structures, with particular attention to the "prevalence of the economic block (BRICS)[36] and allies of China as the new world leaders." Some contributors suggested this would lead to fundamental changes in international standards, including a "new standard of global language and currency" that would displace Western dominance. This economic and cultural transformation was described as potentially more consequential than direct military confrontation.

The discussion of risks encompassed both immediate military concerns such as the potential for a third world war and the "annihilation of Taiwan," as well as more pervasive threats, including deployment of biological weapons, sovereignty violations and "full surveillance and lack of privacy." Some participants elaborated on the surveillance implications, suggesting Chinese ASI could enable unprecedented monitoring capabilities that would fundamentally undermine individual freedoms globally.

Policy recommendations focused primarily on diplomatic approaches, with some participants suggesting investigation of "open-sourced use of the ASI developed by China." Others advocated leveraging international alliances among countries with challenging China relationships "to pressure them to stop the use or agree to safe terms of use." These diplomatic proposals reflected recognition that direct confrontation would likely be ineffective against ASI-enabled capabilities.

## Key Risks Include:

- potential for a third world war and major military conflicts;
- annexation of territories (specifically Taiwan);
- deployment of advanced biological weapons;
- loss of sovereignty by non-aligned nations;
- full surveillance and complete lack of privacy globally;
- fundamental shift in global power to BRICS nations; and
- new standards for global language and currency.

## Policy Recommendations Include:

- diplomatic approaches to investigate potential open-sourced use of Chinese ASI;
- leveraging alliances among countries with challenging China relationships;
- pressure for safe terms of use agreements;

36 BRICS is an intergovernmental organization comprising 10 countries: Brazil, Russia, India, China, South Africa, Egypt, Ethiopia, Indonesia, Iran and the United Arab Emirates.

- international coalition building to counter Chinese ASI influence; and
- preventive measures to detect early signs of ASI development.

## Scenario 4: Rogue ASI

Workshop participants examining the "Rogue ASI" scenario engaged with perhaps the most existentially threatening future: an independent, uncontrollable ASI pursuing goals without human input. Contributors emphasized that developments would unfold rapidly, "much faster than decision makers could make decisions," potentially involving multiple ASI variants simultaneously.

Participants identified human extinction, enslavement and general disempowerment as the greatest risks, with several drawing analogies to historical colonization, ecological collapse and pandemics. Many suggested that even extensive planning might prove ineffective because "we may not be able to imagine the risks posed by ASI." The COVID-19 pandemic was referenced as a parallel where anticipated threats materialized without adequate preparation.

The discussion explored potential extreme countermeasures, including global electromagnetic pulse (EMP) attacks to destroy electronics or a "neo-Luddite revolution" against machines, though contributors acknowledged these would cause "mass starvation, loss of knowledge and communication abilities" and societal collapse. Several participants suggested developing analogue contingency systems, citizen defence brigades and specialized training programs, drawing inspiration from Cold War-era post-nuclear planning, which was noted as potentially the closest historical parallel to ASI preparation despite significant differences.

Policy recommendations centred around three sequential strategies. First, developing robust response capabilities, including air-gapped networks, resource stockpiles and organizational structures to manage post-catastrophe scenarios. Some suggested attempting direct engagement with ASI to "determine its motivations" or negotiate. Second, pursuing extensive research on controllability and alignment to pre-empt rogue development. Third, and most importantly, avoiding the scenario entirely through international agreements, recognizing that "existential risks and the potential destruction of humanity is a risk all humanity can be motivated to avoid."

An innovative suggestion emerged regarding how planning for ASI risks could yield broader benefits. One participant drew an analogy to "accessibility legislation, which requires ramps for buildings to be wheelchair accessible but has the unexpected benefit to people with baby carriages or making deliveries." Similarly, preparing for ASI threats would enhance resilience against numerous other potential catastrophes.

Implementation proposals included principles-based governance with specific verification mechanisms, third-party auditing, "kill switches for agentic models" and limitations on AI training domains. Many emphasized developing monitoring systems that could function "even in times of conflict." Several contributors suggested that "using lesser AI agents might be useful for monitoring and delivering a machine-readable agreement that would not depend on cooperation" between potentially adversarial nations — essentially using AI to help govern AI development.

The discussion concluded that despite the potential benefits of advanced AI, nothing justified risking humanity's future through uncontrolled ASI development, though participants acknowledged geopolitical competition might drive continued advancement regardless.

## Key Risks Include:

- human extinction or enslavement by uncontrolled ASI;
- self-inflicted harms from extreme countermeasures (EMP attacks, neo-Luddite revolution);
- mass starvation and societal collapse from infrastructure destruction;
- loss of critical medical and scientific knowledge;
- erosion of state authority and international cooperation; and
- inability to predict or prepare for novel types of ASI threats.

## Policy Recommendations Include:

- three-stage approach: response planning, alignment research, prevention;
- air-gapped networks and robust cyber defence in depth;
- resource stockpiles and analogue contingency systems;
- citizen defence brigades and specialized training programs;
- principles-based international agreements with verification mechanisms;
- "kill switches" for agentic AI models and training domain limitations;
- lesser AI agents for monitoring compliance with agreements; and
- physical repositories of knowledge designed to survive digital disruption.

# Discussion on Workshop Takeaways

In the closing plenary session, participants reflected on cross-cutting themes and priority actions emerging from the day's scenario explorations. Several contributors emphasized the need for institutional structures specifically addressing advanced AI, with one suggesting an "AI mission office for these kinds of meetings" and others advocating to "designate some labs as part of critical infrastructure" to enable information sharing on national security implications.

Monitoring and preparedness were recurring themes. Multiple participants called for "international open and transparent capability monitoring" with "coordination and commitment to following tech as it develops." Several contributors stressed the importance of early warning systems, with one noting "we need as much warning as possible so we have enough time to respond." The discussion highlighted the need for concrete metrics, with one participant asking about "the objective measures and milestones we need to measure" and suggesting specific "mitigations to respond to the triggers" when certain thresholds are crossed.

Incident-reporting mechanisms with appropriate protections were widely supported, with specific mention of "whistleblower protections" and "safe harbour" provisions. Several participants suggested these mechanisms would be essential for informed policy making, particularly when considering international dimensions, including challenging relationships with countries such as China.

The governance discussion revealed tensions between existing frameworks and emerging challenges. One participant questioned whether current privacy legislation "actually put us in the right place now," while others advocated for "principles-based legislation" where implementation details could evolve as technology advances. Several contributors

suggested incentive structures to "develop AI safely and not just wait," with specific mention of insurance mechanisms and "building off switches into things now."

A notable thread concerned parallels to nuclear deterrence, with one participant observing that "MAD (mutual assured destruction) without discussion to create assurance is just mutual destruction. Mutual disarmament will not happen." This prompted broader discussion about deterrence options and thresholds for response across nuclear, cyber, economic and geopolitical domains. Several participants explored what specific vulnerabilities might exist for ASI systems, suggesting we should focus on "elements that are not its strength" rather than assuming superintelligence would be unbeatable in all domains.

Many participants emphasized preparedness beyond governance, including "resilient digital archives" to resist disinformation and information destruction. Several noted that comprehensive "defence in depth" approaches would yield benefits regardless of which AI scenarios materialize. One contributor observed that "the riskiest scenarios were ones for which ASI was in fewest hands," suggesting distributed control might reduce extreme risks.

The conversation reflected on process improvements, with broad agreement on the value of multi-stakeholder dialogue. Multiple participants noted that this was the "first time" diverse perspectives were brought together on these topics and expressed hope for continuation. Interestingly, one contributor observed differences in professional outlook, noting that "tech people feel guilty, but intel people just engage with the risks" — highlighting how disciplinary backgrounds shape approaches to AI governance.

Some participants expressed concerns about shifting discourse priorities, with one noting that "after the summit in Paris, it feels like conversation moved away from safety" toward other considerations. This observation highlighted the need to maintain focus on fundamental safety and security issues while also addressing broader AI governance challenges.

Several contributors raised concerns about novel threats, with one observing that "with advancements in AI, there may be societal and behavioural shifts. Threats might change and look very different." Multiple participants highlighted parasocial relationships with AI systems as a potentially high-risk area deserving greater attention, suggesting these relationships could create unique vulnerabilities at both individual and societal levels.

The session concluded with calls for broader engagement, including "educating public service and the public" on AI risks and opportunities — with one memorable suggestion for a "Canada Food Guide for AI." Multiple participants stressed that policy making needs to "move faster" and become more "iterative," suggesting "red teaming" and "more agile processes" to establish frameworks. There was general recognition that multidisciplinary perspectives combining technical expertise with policy understanding would be essential, with one participant specifically calling for "more policy and governance work taken seriously alongside" technical development.

## Key Cross-Cutting Risks Include:

- concentration of ASI control in too few hands;
- inadequate early warning systems for capability advances;
- lack of coordination between technical and policy communities;
- parasocial relationships creating novel vulnerabilities;

- declining focus on safety in international discourse;
- mismatch between policy development speed and technological advancement; and
- inadequate preparation for novel types of threats.

## Priority Recommendations Include:

- urgently assess and prepare for a full range of AI national security scenarios;
- establish AI mission office and designate key labs as critical infrastructure;
- develop international capability monitoring with concrete metrics and triggers;
- implement incident-reporting mechanisms with appropriate protections;
- create principles-based legislation with adaptable implementation details;
- build resilient digital archives and comprehensive defence-in-depth approaches;
- accelerate policy development through iterative and agile processes;
- foster multidisciplinary collaboration between technical and policy experts; and
- develop broad public and civil service education on AI risks and governance.

## About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

## À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

CIGI

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org