Implications of Artificial General Intelligence on National and International Security

Yoshua Bengio

As highlighted in the International Scientific Report on the Safety of Advanced AI, the capabilities of general-purpose AI systems have been steadily increasing over the last decade, with a pronounced acceleration in the last few years.[1] If these trends continue, and as per the declared goals of the leading AI companies, we are likely to achieve human-level capabilities across a broad spectrum of cognitive skills, what is commonly called Artificial General Intelligence (AGI). It is remarkable that we have already achieved human-level competence in natural language, i.e., systems that can read and understand texts, and fluently respond or generate new textual, visual, audio or video content. And while scientific advances are impossible to predict precisely, many leading researchers now estimate the timeline to AGI could be as short as a few years or a decade. This is consistent with the steady advances of the last decade driven by algorithmic progress and scaling up the amount of computing resources used, and by the exponential increase in global AI R&D investments well into the trillions of dollars.[2] While the lack of internal deliberation abilities, i.e., thinking, has long been considered one of the main weaknesses of current AI, a recent advance based on a new form of AI with internal deliberation suggests we might potentially be on the brink of bridging the gap to human-level reasoning.[3, 4]

Moreover, frontier AI companies are seeking to develop AI with a specific skill that could very well unlock all others and turbocharge advances: AIs with the ability to advance research in AI. An AI system that would be as capable at AI research as the topmost handful of researchers in an AI Lab would multiply the advanced research workforce by orders of magnitude. Although it takes tens of thousands of GPUs to train the AI, once trained it can be deployed at inference-time in parallel, yielding the equivalent of hundreds of thousands of automated AI workers. Such scaling up could greatly accelerate the path towards superhuman AI systems. The materialization of this scenario could lead to a fast transition from AGI to *Artificial Super-Intelligence* (ASI), ranging from a few months to a few years according to some experts.[5] Imagining such possibilities can be challenging, and we have no guarantee that they will materialize, as the pace and direction of future AI development are largely dependent on the political decisions and scientific advances in months and years ahead. However, given the consequences of some of the scenarios among those outlined by experts as plausible, we now need to seriously consider how to mitigate them.

If ASI arises, what could be the consequences? Clearly, the potential benefits are tremendous and could enable both significant economic growth and great improvements in the well-being of societies, through advances in medicine, education, agriculture, fighting climate change, and more. However, such superior intelligence could also provide unequaled strategic advantages on a global scale and tip the balance in favor of a few (companies, countries or individuals), while causing great harm to many others. This is particularly true in the current geopolitical and corporate contexts whereby control of these technologies is extraordinarily concentrated.[6] Societies would have to address a number of questions: Who will control this great power and to what ends? Could such concentrated power threaten democracy? Beyond the danger of malicious use, do we even have the knowledge and capacity to control machines that are smarter than humans? ASI would open a Pandora's box, enabling both beneficial and destructive outcomes, possibly at the scale of the current existential risks.[7] A significant fraction of AI researchers acknowledge the possibility of such risks: A recent survey of nearly 3000 authors of machine learning papers at recognized scientific venues shows that "between 37.8% and 51.4% of respondents gave at least a 10% chance to advanced AI leading to outcomes as bad as human extinction."[8]

There are also scientific reasons for these concerns (see also the above-cited report for more references).[9, 10] First, one has to remember the basics of AI: the ability to correctly answer questions and achieve goals. Hence, with an ASI, whoever dictates those questions and goals could exploit that intellectual power to effectively have stronger scientific, psychological and planning abilities than other human organizations, and could use that power to enhance their own strength, potentially at the expense of the greater collective.[11] Malicious use of AI could gradually enable extreme concentrations of power, including dominance in economic, political or military terms, if no counter-acting power is in place to prevent any ASI from acquiring a decisive strategic advantage. Second, the ASI could be the controlling entity, if it has as a goal its own preservation. In this case, it would most probably execute subgoals to increase its probability of survival. An ASI with a primary self-preservation goal could notably scatter offspring in insecure computing systems globally, and speak fluently and extremely persuasively in all major languages. If this were to happen before we figure out a way to ensure ASI is either aligned with human interests, or subject to effective human oversight and control, this could lead to catastrophic outcomes and major threats to our collective security. Keeping in mind that some humans would (rightly) want to turn off such a machine, precisely to avoid harm, it would be in the advantage of the AI to (1) try to make sure it is difficult for humans to turn it off, e.g., by copying itself in many places across the internet, (2) try to influence humans in its favor, e.g., via cyberattacks, persuasion, threats, and (3) once it has reduced its dependence on humans (e.g., via robotic manufacturing), aim to eliminate humans altogether, e.g., using a new species-killing virus.

There are many trajectories that could lead to the emergence of an AI with a self-preservation goal. A human operator could specify this goal explicitly (just like we type queries in ChatGPT), for example to advance an ideology (some groups have the stated objective of seeing ASI replace humanity as the dominant entities).[12, 13] But there are also mathematical reasons why self-preservation may emerge unintentionally. By definition, AGI would achieve at least human-level autonomy, if merely given access to a command line on its own servers. In order for an autonomous agent to ensure the highest possible chance of achieving almost any long-term goal, including goals given by human operators, it will need to ensure its preservation.[14] If we are not careful, that implicit self-preservation goal could lead to actions against the well-being of societies, and there are currently no highly reliable techniques to design AI that is guaranteed to be safe.[15] It is also worth noting that given the immense commercial and military value of enabling strong agency in frontier AI systems, (i.e., allowing the AI to not only answer questions but also to plan and act autonomously) there are powerful incentives to invest significant R&D efforts in this pursuit.[16] The current state-of-the-art method seeking to evolve AIs into agents through reinforcement learning techniques involves creating systems that seek maximum positive rewards—thus opening up the possibility of the AI eventually being capable of overtaking the reward system itself.[17, 18]

Major AGI and ASI National Security Challenges

If and when AI systems are able to operate at or above human-level intelligence and autonomy, there would be an unprecedented level of risk for national and international security. Moving towards action to start mitigating these threats is urgent, both because of the unknown timeline for AGI and ASI, and the plausible speed-gap between implementing guardrails, countermeasures and international agreements versus deploying the next frontier AI systems, especially in the current regulatory environment, which imposes little to no restrictions.

It is useful to categorize the different kinds of threats because their mitigations may differ, and we need to find solutions to each that do not worsen the others. In general, there will be many unforeseen effects and the potential for catastrophic outcomes, all calling for caution.

a. National security threats from adversaries using AGI/ASI: Even before the possible emergence of AGI, malicious actors could use future advanced AI to facilitate mass destruction, with threats ranging from CBRN (chemical, biological, radiological, nuclear) to cyber attacks. The recently revealed OpenAI o1 model (September 2024) is the first model to cross the company's own boundary from "low risk" to "medium risk" for CBRN capabilities – the maximum level of risk before OpenAI's policies would preclude releasing a model. Such threats will only increase as AI capabilities continue to rise. Advances in autonomy and agency should be monitored closely, and it should be expected that o1's progress in reasoning abilities and simple solvable planning problems (35.5% to 97.8% on Blocksworld) could soon open the door to better long-term planning and thus improved AI agency.[19] This could yield great economic and geopolitical value but would also pose significant threats to people and infrastructures, unless political and technical guardrails and countermeasures are put in place to prevent AGI systems from falling into the wrong hands.

b. Threats to democratic processes from current and future AI: Deepfakes are already used in political campaigns and to promote dangerous conspiracy theories. The negative impact of future advances in AI could be of a much larger magnitude, and AGI could significantly disrupt societal and power equilibria. In the short term, we are not far from the use of generative AI to design personalized persuasion campaigns. A recent study compared GPT-4 with humans in their ability to change the opinion of a subject (who doesn't know whether they are interacting with a human or a machine) through a text-based dialogue.[20] When GPT-4 has access to the subject's Facebook page, that personalization enables substantially greater persuasion than that from humans. We are only one small step away from greatly increasing such a persuasion ability by improving the persuasion skills of generative models with additional specialized training (called fine-tuning). A state-of-the-art open-source model such as Llama-3.1 could likely be used for such a purpose by a nefarious actor. Such a threat will likely grow as persuasion abilities of advanced AI increase, and we may soon have to face superhuman persuasion abilities given large language models' high competency in languages (already above the human average). If this is combined with advances in planning and agency (to achieve surgical and personalized goals in opinion-shaping), the effect could be highly destabilizing for democratic processes in favor of a rise in totalitarian regimes.[21]

c. Threats to the effective rule of law: Whoever controls the first ASI may gain enough power — through cyberattacks, political influence or enhanced military force — to inhibit other players with the same ASI goal. Such a centralization of power could happen either within or across national territories, and would be a major threat to states' sovereignty. The temptation to use ASI to increase one's power will be strong for some, and may be rationalized by fear of adversaries (including political or business opponents) doing it first. Modern democracy emerged from early information and communication technologies (from postal systems and newspapers to fax machines) where no single human could easily defeat a majority of other humans if they could communicate and coordinate in a group.[22, 23] But this fundamental equilibrium could be toppled with the emergence of the first ASI.

d. Threats to humanity from loss of human control to rogue AIs: Rogue AIs could emerge anywhere (domestically or internationally), either because of carelessness (impelled by pressure from a military arms race, or a commercial equivalent) or intentionally (because of ideological motivations). If their intelligence and ability to act in the world is sufficient, they could gradually control more of their environment, first to immediately protect themselves and then to make sure humans could never turn them off. Note that this threat could be increased by the emergence of more totalitarian regimes which lack good self-correcting mechanisms and might unintentionally allow the emergence of a rogue ASI because of their political agendas.

Government Interventions for Risk Mitigation

The above risks must be mitigated together: Addressing one but not the others could prove to be a monumental mistake. In particular, racing to AGI to land there before adversaries could greatly increase the democratic (b,c) and rogue AI (d) risks. Navigating the required balancing act sufficiently well will be challenging, and the ideas below should be taken as the beginning of a much-needed global effort, one that will require our best minds and substantial investments and innovations, in both science and governance.

1. Major and urgent R&D investments are needed to develop AI with safety guarantees that would continue to be effective if current AIs are scaled to the AGI and ASI levels, with the objective to deliver solutions before AI labs reach AGI.[24] Some of that safety-first frontier AI R&D can be stimulated with regulatory incentives

- such as clarifying liability for developers for systems that could cause major harm - but may ultimately be better served through non-profit organizations with a public protection mission and robust governance mechanisms designed to avoid the conflicts of interest that can emerge with commercial objectives.[25] This is particularly important to address challenges (a), (b) and especially (d), for which no satisfactory solution currently exists. Note that some parts of that R&D, if shared globally, would help mitigate risks (see a,b,c,d), e.g., identifying general means to evaluate risks and how to put technical guardrails, while other parts of that R&D, if shared globally, would increase all those risks, e.g., by increasing capabilities (since improving the prediction of future harm, which is useful to build technical guardrails, requires advances in capabilities).

2. Governments must have substantial visibility and control over these technological advances to avoid the above scenarios, for example to reduce the chances that an ASI developed by an AI lab be deployed without the necessary protections or accessed by a malicious actor. Since frontier AI development is currently fully in private hands, a transition will be required to ensure AGI is developed with national security and other public goods as more central goals, beyond economic returns. Regulation can provide necessary controls, but stronger options may be needed. Outright nationalization, at least in the current environment, is unlikely but promoted by some.[26] Another possibility is that of private-public partnerships, with frontier efforts mostly led under a secured governmental umbrella, with safe commercial spin-offs.[27] This would help in addressing all four challenges (see a,b,c,d) above, especially by securing the most advanced models and imposing appropriate guardrails.[28]

3. This unprecedented context calls for an innovation in the checks-and-balances and multilateral governance mechanisms and treaties for AGI development. To make sure that no single individual, corporation or government could unethically use the power of ASI for their own benefit, up to and including a self-coup, will require institutional and legal changes that could take years to develop, negotiate and implement.[29] Although governments can oversee and regulate their domestic AI labs, international agreements could reduce the possibility that one country creates a rogue AI (see (d) in the above section) or use ASI against populations and infrastructure of another country (a). In this context, what kind of organization with multilateral governance should be developing AGI and ASI? Organizations that are government-funded but at arm's length could play a critical role in balancing interests, countering conflicts of interest and ensuring public good. For example, CERN is a non-governmental entity created by international convention, subject to regulations of its host countries, which are in turn subject to governance obligations to the IAEA, a separate non-governmental entity created by a separate international treaty. A strong design of governance mechanisms (both in terms of rules and of technological support) is crucial, both to avoid abuse of a potential controlled ASI's power (see c), as well as safety compromises arising from race dynamics (between companies or between countries), which could lead to a rogue AI (see d). Furthermore, creating a network of such non-profit multilaterally governed labs located in different countries could further decentralize power and protect global stability.[30]

4. International treaty compliance verification technology will be required sooner than later, but take time to develop and deploy. Past treaties on nuclear weapons became possible thanks to treaty compliance verification procedures and techniques. An AGI treaty would only be effective to prevent a dangerous arms race (or be signed in the first place) if we first develop procedures and technology to verify AI treaty compliance. There currently exists no such reliable verification mechanism, which means that the compliance to any international agreement would be impossible to assess. Software may at first seem hard to govern. However, hardware-enabled governance mechanisms, ideally sufficiently flexible to adapt to changes in governance rules and technology, and the existence of major bottlenecks in the AI hardware supply chain, could enable technological solutions to AI treaty compliance verification.[31, 32, 33]

Given the magnitude of the risks and the potentially catastrophic unknown unknowns, reason should dictate caution and significant efforts to better understand and mitigate those risks, even if we think that the likelihood of catastrophes is small. It will be tempting to accelerate to win the AGI race, but this is a race where everyone could lose. Let us instead use our agency while we still can and deploy our best minds to move forward with sufficient understanding and management of the risks with robust multilateral governance to avoid its perils and thus reap the benefits of AGI for all of humanity.

Acknowledgements: The author thanks David (davidad) Dalrymple, Valérie Pisano, Marc-Antoine Guérard, Susan Rice, Benjamin Prud'homme, Lawrence Lessig, Jason Matheny, and Daniel Privitera for feedback.

Yoshua Bengio is recognized worldwide as one of the leading experts in artificial intelligence, most known for his pioneering work in deep learning which earned him the 2018 A.M. Turing Award, "the Nobel Prize of Computing," with Geoffrey Hinton and Yann LeCun. He is Full Professor at Université de Montréal, and the Founder and Scientific Director of Mila – Quebec AI Institute. He co-directs the CIFAR Learning in Machines & Brains program as Senior Fellow and acts as Scientific Director of IVADO and holds a Canada CIFAR AI Chair. In 2023, he was awarded the prestigious Gerhard Herzberg Canada Gold Medal for Science and Engineering, the country's most significant scientific prize. He is the computer scientist with the highest h-index citation metric and ranked third among all disciplines by scientific impact according to the Stanford bibliometric study. He is a Fellow of both the Royal Society of London and Canada, ACM Fellow, Knight of the Legion of Honor of France, Officer of the Order of Canada, Member of the UN's Scientific Advisory Board for Independent Advice on Breakthroughs in Science and Technology and Scientific Advisor for the UK AI Safety Institute (AISI). Concerned about the social impact of AI, he actively contributed to the Montreal Declaration for the Responsible Development of Artificial Intelligence and devotes himself to reducing the catastrophic risks of future AI.

[1] "International Scientific Report on the Safety of Advanced AI INTERIM REPORT" (Department for Science, Innovation and Technology and AI Safety Institute, May 17, 2024), https://assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

[2] Anson Ho et al., "Algorithmic Progress in Language Models" (arXivLabs, March 9, 2024), https://arxiv.org/pdf/2403.05812.

[3] "Learning to Reason with LLMs" (OpenAI, September 12, 2024), https://openai.com/index/learning-to-reason-with-llms/.

[4] Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati, "LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's O1 on PlanBench" (arXivLabs, September 20, 2024), https://arxiv.org/abs/2409.13373.

[5] Leopold Aschenbrenner, "Situational Awareness: The Decade Ahead," situational-awareness.ai, June 2024, https://situational-awareness.ai/.

[6] "International Scientific Report on the Safety of Advanced AI INTERIM REPORT."

[7] "Statement on AI Risk" (Center for AI Safety, n.d.), https://www.safe.ai/work/statement-on-ai-risk.

[8] Katja Grace et al., "Thousands of AI Authors on the Future of AI" (arXivLabs, January 5, 2024), https://arxiv.org/abs/2401.02843.

[9] Michael Cohen, Marcus Hutter, and Michael Osborne, "Advanced Artificial Agents Intervene in the Provision of Reward," *AI Magazine* 43, no. 3 (August 29, 2022): 282–93, https://doi.org/10.1002/aaai.12064.

[10] "International Scientific Report on the Safety of Advanced AI INTERIM REPORT."

[11] Nazli Choucri and Robert C. North, "Dynamics of International Conflict: Some Policy Implications of Population, Resources, and Technology," *World Politics* 24, no. S1 (1972): 80–122, https://doi.org/10.2307/2010560.

[12] Ellen Huet, "A Cultural Divide over AI Forms in Silicon Valley," Bloomberg.com, December 6, 2023, https://www.bloomberg.com/news/newsletters/2023-12-06/effective-accelerationism-and-beff-jezos-form-new-tech-tribe.

[13] Rich Sutton, "AI Succession," YouTube, September 8, 2023, https://www.youtube.com/watch?v=NgHFMolXs3U.

[14] Stephen Omohundro, "The Basic AI Drives" (Self-Aware Systems), accessed October 16, 2024, https://steveomohundro.com/wp-content/uploads/2009/12/ai_drives_final.pdf.

[15] "International Scientific Report on the Safety of Advanced AI INTERIM REPORT"

[16] Mustafa Suleyman, *The Coming Wave* (Crown, 2023).

[17] Huet, "A Cultural Divide over AI Forms in Silicon Valley."

[18] Cohen, Hutter, and Osborne, "Advanced Artificial Agents Intervene in the Provision of Reward," 282–93.

[19] Valmeekam, Stechly, and Kambhampati, "LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's O1 on PlanBench."

[20] Tanya Petersen, "AI's New Power of Persuasion: It Can Change Your Mind" (EPFL, April 15, 2024), https://actu.epfl.ch/news/ai-s-new-power-of-persuasion-it-can-change-your-mi/.

[21] Yuval Noah Harari, Nexus (Random House, 2024).

[22] J. M. Adelman, "'A Constitutional Conveyance of Intelligence, Public and Private': The Post Office, the Business of Printing, and the American Revolution," *Enterprise and Society* 11, no. 4 (June 24, 2010): 709–52, https://doi.org/10.1093/es/khq079.

[23] Michael Curtis, "Catalyzing the Collapse: The Computer and the Fall of the Soviet Union" (Disciplinary Deliverable, n.d.), https://phoenixfiles.olin.edu/do/1163/iiif/2a41f3e8-4562-44ae-8505-56c91b17cda0/full/full/0/mcurtis06_ahs_final.pdf.

[24] Yoshua Bengio et al., "Can a Bayesian Oracle Prevent Harm from an Agent?" (arXivLabs, August 22, 2024), https://arxiv.org/pdf/2408.05284.

[25] Yoshua Bengio, "AI and Catastrophic Risk" (*Journal of Democracy,* September 2023), https://www.journalofdemocracy.org/ai-and-catastrophic-risk/.

[26] Charles Jennings, "There's Only One Way to Control AI: Nationalization" (POLITICO, August 20, 2023), https://www.politico.com/news/magazine/2023/08/20/its-time-to-nationalize-ai-00111862.

[27] David Ignatius, "No Manhattan Project for AI, but Maybe a Los Alamos" (*The Washington Post*, September 6, 2024), https://www.washingtonpost.com/opinions/2024/09/06/general-artificial-intelligence-biden-administration-technology-strategy/.

[28] Sella Nevo et al., "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models" (RAND, May 30, 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html.

[29] "Self-Coup," Wikipedia (Wikimedia Foundation, October 13, 2024), https://en.wikipedia.org/wiki/Self-coup.

[30] Yoshua Bengio, "AI and Catastrophic Risk."

[31] Gabriel Kulp et al., "Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified under Export Control Classification Numbers 3A090 and 4A090" (RAND, January 18, 2024), https://www.rand.org/pubs/working_papers/WRA3056-1.html.

[32] James Petrie et al., "Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees," 2024, https://yoshuabengio.org/wp-content/uploads/2024/09/FlexHEG-Interim-Report_2024.pdf.

[33] Onni Aarne, Tim Fist, and Caleb Withers, "Secure, Governable Chips" (Center for a New American Security, January 8, 2024), https://www.cnas.org/publications/reports/secure-governable-chips.