sipri

ADDRESSING THE RISKS THAT CIVILIAN AI POSES TO INTERNATIONAL PEACE AND SECURITY

The Role of Responsible Innovation

VINCENT BOULANIN, JULES PALAYER AND CHARLES OVINK

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE

SIPRI is an independent international institute dedicated to research into conflict, armaments, arms control and disarmament. Established in 1966, SIPRI provides data, analysis and recommendations, based on open sources, to policymakers, researchers, media and the interested public.

The Governing Board is not responsible for the views expressed in the publications of the Institute.

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE

Signalistgatan 9
SE-169 70 Solna, Sweden
Telephone: +46 8 655 97 00
Email: sipri@sipri.org
Internet: www.sipri.org

ADDRESSING THE RISKS THAT CIVILIAN AI POSES TO INTERNATIONAL PEACE AND SECURITY

The Role of Responsible Innovation

VINCENT BOULANIN, JULES PALAYER AND CHARLES OVINK

November 2025

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE

United Nations Office
for Disarmament Affairs

Funded by
the European Union

DOI: https://doi.org/10.55163/OXHH2798

Contents

Acknowledgements

This report is part of a joint project by the Stockholm International Peace Research Institute and the United Nations Office for Disarmament Affairs on Promoting Responsible Innovation in Artificial Intelligence for Peace and Security, which is funded by a decision of the Council of the European Union (EU) (Council Decision (CFSP) 2022/2269 of 18 November 2022). The authors express their sincere gratitude to the EU for the generous financial support for the project.

The authors are also indebted to the more than 200 experts from academia, industry, governments and international organizations who shared their knowledge and views at the different stages of the project. The authors are particularly grateful for the insights provided by the members of the project's advisory board: Raja Chatila, Frank Dignum, Edson Prestes, Ludovic Righetti and Julia Stoyanovich and the late Abhishek Gupta.

The authors also thank colleagues who provided support throughout the project and the writing of this report: Gaston Collazo, Beyza Unal, Katherine Prizeman, Laura Bruun, Alexander Blanchard, Leonard Günzel, Vivianne Manlai, Rafael Lipcsey, Kolja Brockmann and Amelie Lutz. Finally, the authors would like to acknowledge the invaluable editorial work of Linda Nix and the SIPRI Editorial Department.

The views and opinions in this report are solely those of the authors and do not represent the official view of SIPRI, UNODA or the EU. Responsibility for the information set out in this report lies entirely with the authors.

Executive summary

This report provides an overview of how advances in artificial intelligence (AI) in the civilian domain could present risks to international peace and security, and how such risks can be addressed through responsible innovation. The report's key findings can be summarized as follows.

Advances in civilian AI impact international peace and security in multiple ways. AI systems can be misused for influence operations, cyberattacks and developing weapons systems. They can also inadvertently reinforce trends that undermine the foundations of sustainable peace and security. Generative AI, for example, has been accelerating the pollution of the information ecosystem and contributing to the erosion of trust in public discourse and political institutions. Open-source release of highly capable general-purpose AI models could impact how countries compete in the development of AI-enabled military technologies. While expert views on the likelihood and severity of these scenarios diverge, most of these risks could be prevented or mitigated through the systematic use of responsible innovation practices within the AI community. When properly employed, the set of practices responsible innovation involves can help AI practitioners and companies to identify risks associated with their research or product, including risks outside their immediate frame of reference, and to adopt measures that can meaningfully reduce the likelihood or scale of a given risk's impact.

Responsible innovation practices within the AI community are progressing, including in relation to issues directly connected to international peace and security. Companies developing and deploying the most advanced AI models routinely deploy technical and procedural measures to minimize the likelihood and potential impacts of their model's misuse for political, criminal and violent purposes. There is also a lively conversation among AI experts about practices that can make AI safer, more secure, more trustworthy and, ultimately, less likely to generate large-scale harm. However, that progress has been uneven across the AI industry and academia. For instance, small and medium-sized enterprises (SMEs) consulted as part of this project commonly reported that dual-use concerns were rarely a top priority and that they felt far less equipped than major AI companies to integrate and engage with these concerns in their workflows. Similarly, major AI companies have been inconsistent in their efforts and unevenly transparent about their risk management methods. In academia, efforts to mainstream education and capacity building in responsible AI remain limited. Supporting materials from civil society are predominantly in English and rarely link AI innovation explicitly to international peace and security, save for issues like weapons risk.

Self-governance within the AI community will not be sufficient to ensure international peace and security risks associated with civilian AI are identified and addressed in a timely and effective way. AI risk management is a collective action problem that needs governmental interventions and international coordination to level the playing field and ensure that minimum standards are applied across the AI community.

Based on these findings, the report makes the following recommendations to academia, industry, states and international organizations.

To academia: promote responsible innovation practices through exemplarity, education and peer review. Senior academics could lead by example and follow best practices for responsible research and innovation in their work. They can also support the promotion of responsible innovation practices in AI-related curriculums and create opportunities for students to reflect on the societal impact of their future work in technical classes or through peer review. Other actions include fostering interdisciplinary approaches to research, which are key to responsible innovation.

To industry: strengthen responsible innovation practices in the development and deployment of AI products and services. Major AI companies could be more transparent about the methods and processes they use to assess risks; create safer conditions for independent third-party risk evaluation; and more actively support the development of better methods to evaluate risks associated with AI systems based on large language models. SMEs and other entities that develop and deploy specialized AI systems could make greater use of the growing body of online resources on how to engage in responsible innovation, and call on third-party actors to help them assess the spectrum of risks associated with their AI models. Industry organizations could more actively support the emergence and information sharing of best practices for risk management, within and across the industry and academia. Industry organizations could also facilitate the transfer of lessons from fields of science and technology dedicated to safety-critical systems, and provide material resources that are tailored to the needs of SMEs.

To states and international organizations: work together to create the conditions for more universal and consistent adoption of responsible innovation practices that benefit international peace and security. Governments could support education in responsible AI; create or further develop infrastructure and resources for independent testing and evaluations at the national level; and provide resources for companies, especially SMEs, on how to implement responsible innovation practices. International organizations could raise awareness and build a greater common understanding among states on the risks that advances in AI pose to international peace and security; better coordinate among themselves so the efforts to address the risks are not restricted by institutional silos; and facilitate international dialogue, coordination and pooling of resources on safety and security testing and evaluation of AI systems.

1. Introduction

For a long time, the conversation in academic and policy circles on artificial intelligence (AI) and international peace and security was almost exclusively focused on the risks associated with its military applications, such as autonomous weapons systems. Over the past three years, however, a series of events has brought greater attention to the international peace and security risks that could stem from advances in civilian AI. The breakthrough in generative AI, and the reactions that ensued, was perhaps the most significant of these events.[1] It triggered a much-needed discussion in mainstream media and expert circles on how the civilian AI could be misused for harmful purposes, such as large-scale influence operations and sophisticated cyberattacks.[2] It also shed light on how the rapid development and deployment of highly capable general-purpose AI systems could further destabilize the current world order by causing greater inequalities within and between states and by intensifying strategic competition between major military powers.[3]

Although the link between civilian advances in AI and international peace and security is increasingly recognized, it has not been explored systematically. Existing scholarly work often tends to focus on specific risk scenarios (e.g. misuse of AI for influence operations or bioterrorism), technologies (e.g. generative AI or commercial robotics) or security context (e.g. geostrategic competition between the United States and China).[4] The conversation within policy circles is also fragmented along topic lines that reflect the political limitations or priorities of the institutions or forums that host these discussions. For instance, the question of how civilian advances in AI could be weaponized was largely excluded from diplomatic talks that led to the adoption of the Global Digital Compact by the United Nations General Assembly because issues related to the military use of AI were beyond its scope. Meanwhile, the debates at the UN Convention on Certain Conventional Weapons (CCW) on autonomous weapon systems and at summits on the Responsible Use of AI in the Military Domain (REAIM), which has been the epicentre of diplomatic talks on AI and international peace and security for years, have remained laser-focused on the development of AI in the military domain.

This report attempts to bridge that gap in the literature. It aims to present, in a systematic manner, how advances in AI in the civilian domain present risks to international peace and security, and how such risks have been or could be addressed by the civilian AI community—individuals and organizations from the private sector, academia and civil society involved in the AI lifecycle (particularly researching, developing and deploying AI) in the civilian domain—through responsible innovation.

This report is targeted at governmental and non-governmental actors (e.g. from academia, civil society and industry) who contribute to the governance of AI at the international level, particularly those that contribute to policy processes that pertain directly or indirectly to international peace and security. These include the UN Gen-

[1] Future of Life, 'Pause giant AI experiment: An open letter', 22 Mar. 2023; Milmo, D., ' "Godfather of AI" shortens odds of the technology wiping out humanity over next 30 years', *The Guardian*, 27 Dec. 2024; and Hanna, A. and Bender, E., 'AI causes real harm. Let's focus on that over the end-of-humanity hype', *Scientific American*, 12 Aug. 2023.

[2] Brundage, M. et al., *The Malicious Use of Artificial Intelligence: Forecast, Prevention and Mitigation* (Future of Humanity Institute et al.: Oxford, Feb. 2018).

[3] Bengio, Y. et al., *International AI Safety Report*, British Department for Science, Innovation and Technology (DSIT) Research Series No. 2025/001 (AI Action Summit: London, Jan. 2025).

[4] Ekins, S., et al., 'There's a "ChatGPT" for biology. What could go wrong?', *Bulletin of the Atomic Scientists*, 24 Mar. 2023; Goldstein, J. A. et al., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations* (ArXiv: Jan. 2023); Weidinger, L. et al., *Ethical and Social Risks of Harm from Language Models* (Google DeepMind: London, 2021); and Drexel, B., *Promethean Rivalry: The World-altering Stakes in Sino–American AI Competition* (Center for New American Security: Washington, DC, 2025).

eral Assembly First Committee, REAIM, the summits organized by the International Telecommunication Union (ITU) as part of its AI for Good initiative, and the newly created UN Independent International Scientific Panel on AI. The report also aims to be relevant for AI practitioners and AI organizations that seek to better understand how they can prevent their research or products from having negative downstream implications for international peace and security.

The report builds on desk research and insights gathered from a series of events that the authors organized between 2022 and 2025, including seven thematic multi-stakeholder dialogues, two public panel discussions, six closed-door private sector roundtables, five in-person workshops targeted at graduate students in AI and related disciplines, and four roundtables with professors and faculty from science, technology, engineering and mathematics (STEM) faculties. The findings and recommendations are the views of the authors and do not necessarily reflect the views of the individual experts and organizations with which the authors engaged; nor do they reflect the institutional views of the Stockholm International Peace Research Institute (SIPRI), the UN Office for Disarmament Affairs (UNODA) or the funder, the European Union (EU).

The report is structured as follows. Chapter 2 maps the spectrum of risks that advances in civilian AI present to international peace and security, discusses how these risks have been considered in expert and policy conversations, and argues why responsible innovation is a useful approach to address them. Chapter 3 reports on the state of responsible innovation practice in the AI community, focusing on what industry actors and academia have done to identify, evaluate and address risks that advances in AI could pose to international peace and security. It presents the types of measures and practices that the AI community has deployed, and discusses whether these are widespread, which have been successful so far, and where gaps remain. Chapter 4 discusses the need for policy interventions and identifies actions and approaches that international organizations and governments could take to further support responsible innovation practices in the AI community. Chapters 5 and 6 present the report's key findings and recommendations, respectively.

2. Mapping the risks that civilian AI poses to international peace and security

Advances in civilian AI could undermine international peace and security in many ways—some direct, others indirect; some known and well understood, others more speculative. This chapter aims to provide a systematic overview of the types of risks that AI presents to international peace and security, as well as a sense of how these risks have been perceived so far within the AI expert and policy communities. It concludes by noting that, although there is not necessarily consensus on which types of risks demand the most urgent attention, there is a general recognition that greater engagement of the civilian AI community with responsible innovation practices can provide a foundation for the mitigation of all types of risks.

Identifying the risks

The risks that AI can present to international peace and security can be mapped out in various ways. Four crucial dimensions of AI risks (figure 2.1) are:

1. **The causal pathway**—that is, how AI could have negative consequences. A frequently used distinction in this context is between misuse risks, accidental risks and structural risks.[5] *Misuse risk* refers to the possibility that some actors would intentionally use a technology in a way the creators of that technology did not intend and would find undesirable or harmful. *Accidental risk* refers to the possibility that an AI's intended use could have unintended harmful effects. *Structural risks* acknowledge that AI development and deployment can have unintended negative consequences on society at large.
2. **The characteristics of AI**—for example, whether the risks are related to general-purpose models or more specialized applications. In some cases, technical characteristics of the models will play a determining role in the kind of risk they create. For example, depending on the data a model is trained on, or whether it is a generative model, the computational resources and infrastructure needed to retrain or repurpose a model vary.
3. **The domains of impact**—including the political domain (e.g. impact on electoral processes), the cyber domain (e.g. impact on confidentiality or integrity of networked systems and information), and the physical domain (e.g. impact on human life and critical infrastructure).
4. **The state of knowledge**—that is, whether the risks are known or speculative.

The following sections elaborate on the first dimension of AI risk—the causal pathway—as it is a useful lens to unpack the link to international peace and security. It also provides a valuable entry point to evaluate AI risk across the other dimensions.

Misuse risks

The first, and most obvious, way civilian AI could present risks to international peace and security is through cases of misuse—that is, situations where an actor would

[5] Zwetsloot, R. and Dafoe, A., 'Thinking about risks from AI: Accidents, misuse and structure', *Lawfare*, 11 Feb. 2019.

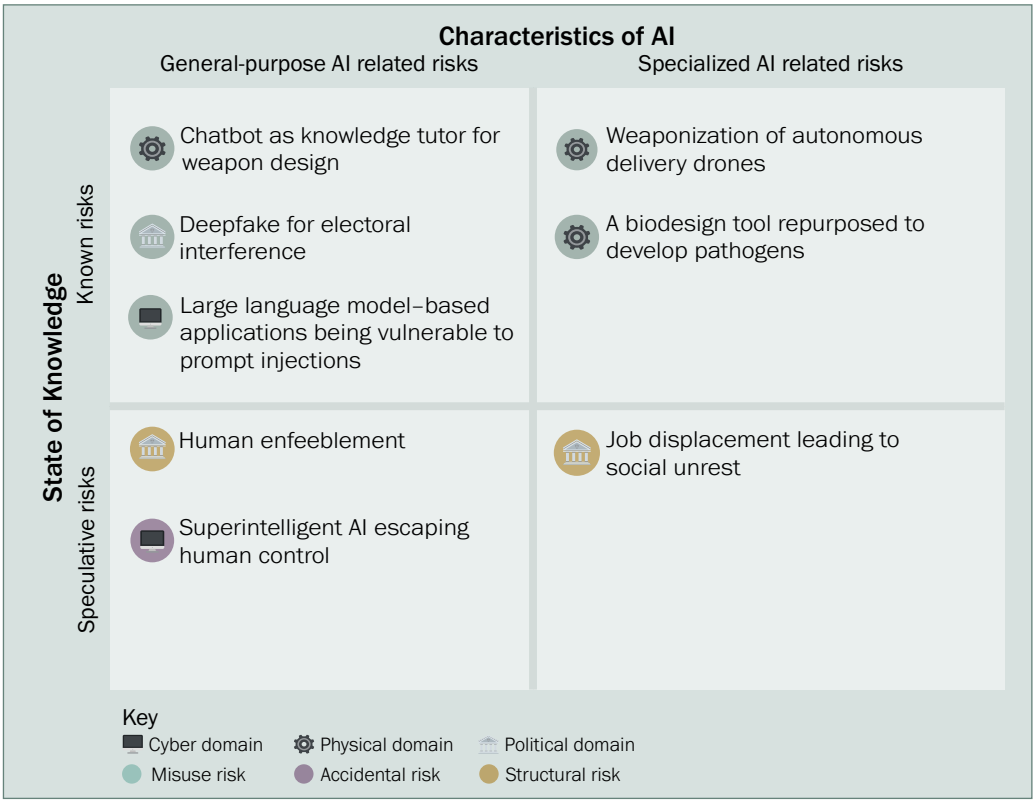

Figure 2.1. Four crucial dimensions of risks that artificial intelligence (AI) presents to international peace and security

intentionally use a technology in a way the creators of that technology did not intend and would find undesirable or harmful.[6] AI is inherently a *dual-use technology* (see box 2.1). Many, if not most, AI applications intended for peaceful purposes can be repurposed or used for harmful ends. Recent history provides numerous examples of AI technologies that have been misused by some state and non-state actors to pursue goals—whether political, economic or military—which had, or could have had, negative implications for international peace and security. Here are three examples that materialized in the political, cyber and physical domains:

Political domain. A precisely timed deepfake was used in the very last days of the Slovak parliamentary election in 2023. Right after the election had entered a 'silent period'—a period during which parties are meant to refrain from discussing election-related matters in the media—an AI-generated audio file surfaced online in which the pro-European candidate appeared to discuss how to conduct electoral fraud.[7] The candidate quickly denied its authenticity, but the clip still went viral, and he eventually lost to a Russia-friendly candidate. Whether that viral diffusion of that clip decided the fate of the election remains debated.[8] The 'Slovak case' was, nonetheless, seen as a blueprint of how AI could be misused to interfere with democratic elections and to serve broader geopolitical objectives or interests (in this case Russian interests, although Russia's direct involvement in the production of the deepfake was never firmly established).

Cyber domain. Anthropic, a major US AI company, reported in August 2025 that its model Claude had been misused by cybercriminals to commit 'large-scale theft and

[6] Brundage et al., *The Malicious Use of Artificial Intelligence: Forecast, Prevention and Mitigation* (note 2).
[7] Meaker, M., 'Slovakia's election deepfakes show AI is a danger to democracy', *Wired*, 3 Oct. 2023.
[8] De Nadal, L. and Jančárik, P., 'Beyond the deepfake hype: AI, democracy, and "the Slovak case"', *Harvard Kennedy School Misinformation Review*, 22 Aug. 2024.

Box 2.1. Defining dual-use

The concept of dual-use was originally coined in the 1990s by the US Office of Technology Assessment to highlight that technologies underlying the development of weapons of mass destruction also had civilian and peaceful purposes. Over time, the use of the term has evolved. In arms control and export control circles, it commonly refers to technology that has both civilian and military uses. In discussions about the societal impact of technology, the term is often used in an even broader sense to convey the general idea that technology may have 'an intended use or primary purpose which is good (or at least not bad) and a secondary purpose or use which is potentially harmful and is not intended by those who developed the technology in the first place'. The broader framing is intended to encompass a more expansive series of uses than military activity, including malicious cyber activities and political influence operations.

Source: Forge, J., 'A note on the definition of dual use', *Science and Engineering Ethics*, vol. 16 (2010).

extortion of personal data'.[9] The cybercriminals used Claude code to plan and execute parts of ransomware operations, including autonomously selecting data to steal, setting ransom amounts and generating threatening ransom notes. This use of AI in the cyber domain is particularly concerning since it boosts the capabilities of actors that would have otherwise needed technical and human resources to conduct the same operations.

Physical domain. Civilian drones developed for recreational or commercial purposes, like farming and professional photography, increasingly include AI-enabled features that allow the drone to navigate autonomously and to identify and track people and objects. States and non-state actors engaged in armed conflicts have been very quick to leverage these features for military operations, even when the companies that develop AI-enabled drones are opposed to their weaponization or military end-use.[10] For instance, the leading actor in the commercial drone market, DJI, has issued an official statement asserting its unequivocal opposition to the weaponization of its products, yet its drones have been used on a massive scale in many conflicts, not least the war in Ukraine.[11]

Admittedly, many technologies can be repurposed for military use or violent use, from machetes to cars. In fact, 10 years ago Toyota, the car company, faced a situation not too different from what DJI experiences today, in that its pick-up trucks had become the vehicles of choice of the terrorist organization ISIS.[12] However, there are three aspects worth noting regarding AI's dual-use potential:

1. AI has significantly lowered the barrier to entry for certain forms of sophisticated and potentially harmful actions.[13] For instance, targeted or large-scale influence operations and cyberattacks are no longer reserved for resourceful or technically skilled actors.[14] General-purpose AI based on large language models (LLMs) has made the production of deepfakes such as the one involved in the Slovak case accessible to virtually any actor who has the time and willingness to spread harmful content. LLMs have also made it possible to conduct a cyberattack without the need to

[9] Anthropic, 'Detecting and countering misuse of AI', Threat Intelligence Report, 27 Aug. 2025.

[10] Bondar, K., *Ukraine's Future Vision and Current Capabilities for Waging AI-enabled Autonomous Warfare* (Center for Strategic & International Studies (CSIS) Wadhwani AI Center: Washington, DC, 2025).

[11] DJI, 'DJI has always opposed combat use of civilian drones and is not a "Chinese military company" ', Media release, 11 Nov. 2022.

[12] 'US officials ask how ISIS got so many Toyota trucks', ABC News, 6 Oct. 2015.

[13] Marchal, N. et al., *Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-world Data* (Google DeepMind: London, 2024).

[14] Weidinger, L. et al, *Sociotechnical Safety Evaluation of Generative AI Systems* (Google DeepMind: London, 2023).

write or manipulate computer code.[15] For instance, AI virtual assistants can be tricked into executing tasks such as leaking private information via so-called 'indirect prompt injections' written in natural language.[16]

2. AI enables malicious actors to operate at a much larger scale. For example, generative AI systems allow politically motivated actors to produce more content, faster, while AI chatbots allow cyber threat actors to acquire key information on potential targets more quickly.[17]
3. AI might introduce novel ways to inflict harm. For instance, one experiment demonstrated how generative AI could be leveraged to design new toxic molecules.[18]

These three aspects and civilian AI's general and enabling nature make its misuse particularly versatile and its impact potentially wide-ranging.

Accidental risks

The risks civilian AI poses to international peace and security could also arise by accident, that is, not involving a malicious or motivated actor purposefully misusing the technology. This is because the intended use of an AI, like any complex technology, can have unintended effects.[19] Accidental harms to peace and security can manifest through technological failure to perform as intended or predicted, or as an unintended consequence of the technology working as it should.[20]

Most AI systems today, including the general-purpose models powered by LLMs, are based on machine learning, an approach to software programming that relies on data and statistical methods. This approach has led to significant increases in AI capabilities, but it has limitations. Machine-learning-based systems can be opaque in their functioning and can fail in unpredictable ways. The consequences of such unpredictability depend on the domain of application and can vary from benign accidents (e.g. a chatbot making erroneous correlations in data that results in a nonsensical or inaccurate output, commonly called 'hallucinations') to lethal accidents (e.g. a self-driving car crashing and killing its passengers).[21] For most civilian applications of AI, technical failures will, most often, have no immediate impact on peace and security. However, in some domains, notably that of information, AI performance failures can compound in problematic ways.

One scenario on the minds of many experts is that the accumulation of inaccuracies generated by AI chatbots on the internet is rapidly polluting the global information ecosystem.[22] This deterioration has important political implications, not least for the health of democracies. It is making it harder for citizens and policymakers not only to discern true information but also to discuss and communicate political decisions. Admittedly, this epistemic problem is already a reality, but it could become more acute

[15] Google Threat Intelligence Group, 'Adversarial misuse of generative AI', Google Threat Intelligence Blog, 29 Jan. 2025.

[16] Heikkilä, M., 'Three ways AI chatbots are a security disaster', *MIT Technology Review*, 3 Apr. 2023.

[17] Google Threat Intelligence Group (note 15).

[18] Urbina, F. et al., 'Dual use of artificial-intelligence-powered drug discovery', *Nature Machine Intelligence*, vol. 4 (2022).

[19] Perrow, C., *Normal Accidents: Living with High-Risk Technologies* (Princeton University Press: Princeton, NJ, 1999).

[20] Grunwald, A., *Technology Assessment in Practice and in Theory* (Routledge: London, 2019).

[21] On AI hallucinations see IBM, 'What are AI hallucinations?', IBM Think Blog [n.d.].

[22] Nirvonen, N. et al., 'Artificial intelligence in the information ecosystem: Affordances for everyday information seeking', *Journal of the Association for Information Science and Technology*, vol. 75, no. 10 (2023).

as AI-generated information becomes more pervasive in society.[23] It also has tangible downstream consequences for international peace and security. Lack of trust in governmental information can be problematic in times of crisis. For instance, should a cyber incident occur at a scale to paralyse a large part of society or the economy, it could become harder for governments to manage reaction from public opinion (e.g. assigning blame to a specific state) and escalatory political discourse that could ensue (e.g. calling for action against an alleged attacker).[24]

Accidental risks are not just about the downstream effects of technical failures. A technology can also have unintended negative consequences even when it is working as it should. An example is the internal combustion engine, which worked so well that it became ubiquitous, but the high levels of emissions that ensued from the engine's rapid and massive adoption have fuelled global warming. A more recent example in the digital realm is social media's recommender algorithms. These were initially designed to provide users with individually curated content. A side effect of their efficacy is that they create echo chambers and filter bubbles, which, among other harmful effects, exacerbate political polarization.[25] In the case of AI, a big question is whether, and if so, how, making AI 'more intelligent' could end up being a risk to humanity. One hypothetical—and highly debated—scenario is that AI systems could become so intelligent that they could escape human control.[26] The possibility that AI could outsmart humans has been contemplated since the dawn of AI as a scientific discipline, including by Alan Turing.[27] The 'loss of control' scenario associated with AI systems that are 'superintelligent' (to use a term of art in the AI community) has also been the focus of numerous science fiction books and movies. However, up until very recently, this scenario was not considered a realistic possibility. That said, the breakthrough in AI capabilities over the past few years changed the equation. Several prominent figures in the AI community, such as Nobel Laureate Geoffrey Hinton, warned that this hypothetical scenario is becoming increasingly plausible.[28] AI systems, as they become more capable or superintelligent could inadvertently develop some 'control-undermining' capabilities, making them able to evade or undermine human control.[29]

The fear is that AI systems could make use of such capabilities in situations where humans would want to limit or shut them down, for instance, if their behaviour was not sufficiently aligned with human interest and values.[30] According to concerned AI safety experts, a concrete illustration would be where a general-purpose AI would take control of critical infrastructure to blackmail human decision-makers into not shutting it down and granting it more power (i.e. control more physical resources).[31] Another possibility that has been considered is that the AI would deceptively manipulate people's opinions to ensure that decision-makers (within government or within industry) make choices that serve the AI's goals.[32] This scenario would have implications for international peace

[23] O'Callaghan, C., 'Postpolitics and post-truth', ed. A. Kobayashi, *International Encyclopaedia of Human Geography*, 2nd edn (Elsevier: Amsterdam, 2020); and Ross, A. et al., 'Echo chambers, filter bubbles, and polarisation: A literature review', Reuters Institute for the Study of Journalism, University of Oxford, 19 Jan. 2022.

[24] Turell, J., Su, F. and Boulanin V., 'Cyber-incident management: Identifying and dealing with the risk of escalation', SIPRI Policy Paper No. 55, Sep. 2020.

[25] Ross et al. (note 23).

[26] Marcus, G., 'AI armageddon? Assessing the threat from artificial general intelligence', *Times Literary Supplement*, 19 Sep. 2025; Yudkowsky, E. and Soares, N., *If Anyone Builds It, Everyone Dies: The Case Against Superintelligent AI* (Bodley Head: London, 2025); and Hanna and Bender (note 1).

[27] Turing, A. M., 'Intelligent machinery, a heretical theory', *Philosophia Mathematica*, vol. 4, no. 3 (1996).

[28] Milmo (note 1).

[29] On loss of control see Bengio et al. (note 3), sect. 2.2.3.

[30] Russell, S., *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin: London, 2019).

[31] Fenwick, C. and Qureshi, Z., 'Risks from power-seeking AI systems', 80,000 Hours, 17 July 2025.

[32] Hendrycks, D., Mazeika, M. and Woodside, T., 'An overview of catastrophic AI risks', *arXiv* 2306.12001v6 (2023).

Box 2.2. Artificial general intelligence

The term 'artificial general intelligence' (AGI) is typically used to describe artificial intelligence (AI) systems that will possess flexible and general intelligence comparable to that of humans. AGI is typically contrasted with artificial 'narrow' intelligence, where an AI system can only excel at discrete 'intelligent' tasks such as recognizing objects or people from images, translating language or playing games. Whether and when AGI has or could be achieved is a matter of great debate in the AI community, in part because AI experts view its defining characteristics differently. For some, the key metric is the *output* of intelligence: what an AI system can do—AGI will be achieved when AI can autonomously and reliably perform any tasks that humans do. According to that view, AGI could be achieved within the next few years (if not already). For others, the key metric is the *process* of intelligence: AGI will be achieved when AI can demonstrate a human-like ability to generalize from experience to very different situations and solve novel problems without prior preparation. According to that view, AGI is still a long way off.

Source: Chollet, F., 'How we get to AGI', Speech, AI Startup School, San Franciso, 16 June 2025; Markus, G., 'AGI versus "broad, shallow intelligence" ', Marcus on AI Blog, 14 Jan. 2025; and Agüera, B. and Norvig, P., 'Artificial general intelligence is already here', *Noéma*, 10 Oct. 2023.

and security: under certain circumstances, the loss of control could lead to disruption or physical harm on a global scale. This is why people and organizations that are concerned by the loss of control scenario have been arguing for greater collaboration between big powers (not least the USA and China) on AI safety. In their view, states that are leading development in AI have a common interest in ensuring that powerful AI systems do not go rogue.[33] It should be noted, however, that the loss of control scenario associated with superintelligent AI systems remains highly contested. Many actors within the AI community find it highly implausible and consider that it unhelpfully diverts attention from more urgent risks, including structural risks.

Structural risks

The misuse/accidental risk dichotomy, which is common in the AI risk literature, tends to focus on the end of the causal chain: how the use of a specific AI or type of AI could purposefully or inadvertently cause harm.[34] However, the progress of AI in general also shapes economic, political and societal structures in ways that could be disruptive or harmful. Here are four frequently discussed scenarios on the broader impact of AI that have relevance for international peace and security.

The first is that substantial increases in the capability of general-purpose AI could reshuffle the current distribution of power within and between states. They could accelerate the process of job automation (which is already underway with present-day AI) and lead to major changes in the global supply chain (e.g. via relocation of production lines).[35] Such changes would have societal and political implications. A rapid rise in unemployment could lead to, among other things, social unrest, the rise of more antagonistic political forces and a potential increase in criminal activities.[36]

At the same time, the countries that are best positioned to leverage advanced AI could enjoy a drastic acceleration in scientific and technological progress, thanks in part to increasing automation of research and development. Such countries would also likely use these capabilities to develop or modernize their military arsenals. This is the reason some pundits see the quest for artificial general intelligence (AGI) (see box 2.2)

[33] Tegmark, M., X, 15 Oct. 2024 <https://x.com/tegmark/status/1846171455899242771>; and Tegmark, M., 'The hopium wars: The AGI entente delusion', *LessWrong*, 12 Oct. 2024.

[34] Zwetsloot and Dafoe (note 5).

[35] On labour market risks see Bengio et al. (note 3), sect. 2.3.1.

[36] Bryson, J., 'What are people for? Employment and the real existential threat of AI', Adventures in NI Blog, 13 Apr. 2020.

as a central component of competition for great power. In 2024 Google's former CEO, Eric Schmidt, together with Dan Hendrycks, an AI safety researcher, and Alexandr Wang, an AI entrepreneur now Chief AI Officer at Meta, published a 'Superintelligence strategy' in which they compared the destabilizing power of highly capable AI to that of nuclear weapons.[37] However, their claim is highly debated. For many actors in the AI community, focusing on risks posed by superintelligent AI systems unproductively diverts attention and resources from more imminent and tangible risks, such as the use of present-day AI to scale up disinformation, or how AI's resource consumption could disrupt the environment and people's livelihood in parts of the world, which can lead to societal unrest and political instability. For more on this debate, see the following section on evaluating risks.

A second scenario is that advances in AI capabilities could lead to 'human enfeeblement', a situation where humans become overly dependent on AI systems.[38] The fact that AI can increasingly perform complex tasks could lead to a situation where humans rely on AI for cognitive tasks, and gradually lose their skills and ability to exercise the moral and critical judgements that are important for human society and political systems. This process creates, in turn, systemic vulnerabilities that state and non-state actors can exploit for political purposes and cognitive warfare.[39] The more people rely on AI systems, the more they become likely to follow suggestions and accept advice that serves the interests of a given actor.[40]

The structural risks of AI for peace and security are not confined to its technical capabilities alone; they are fundamentally embedded in how AI is developed and deployed. A third scenario, in this context, relates to the globalized, resource-intensive nature of AI development, which not only creates significant environmental and social costs that have security implications but also introduces vulnerabilities.

Today's advances in AI require significant volumes of resources: critical minerals for AI hardware, energy and water for data centres, and human labour.[41] The extraction and processing of such resources disproportionately affect countries in the Global South, where limited environmental safeguards and poor labour protections often exist. For example, critical minerals like lithium and cobalt are primarily sourced via practices harmful to local ecosystems and worker health. Similarly, AI companies rely on a huge, globally distributed workforce for data labelling, often subjecting workers in the Global South to poor working conditions, low pay and exposure to graphic content.[42] Such environmental and social consequences have implications for peace and security because they can fuel, among other things, social unrest, political instability and conflict over resources (e.g. around access to water, land or critical minerals).[43]

Furthermore, AI's reliance on a complex and often opaque global supply chain introduces critical security vulnerabilities. Using third-party firms in critical steps in the AI lifecycle, such as data labelling, can enlarge the potential attack surface for malicious actors to exploit models' vulnerabilities or inject malicious data. The provenance and characteristics of the training data also introduce security risks. Over-reliance on datasets centred on English and other high-resource languages also creates a

[37] Hendrycks, D., Schmidt, E. and Wang, A., 'Superintelligence strategy', *arXiv*, 2503.05628v2, 14 Apr. 2025.

[38] Grey, M., and Segerie, C-R., 'The AI risk spectrum: From dangerous capabilities to existential threats', *arXiv* 2508.13700v1, 20 Aug. 2025, Aug. 2025, pp. 72–73.

[39] North Atlantic Treaty Organization, Allied Command Transformation, 'Cognitive warfare', [n.d.].

[40] Rainie, L. and Anderson, J., 'Theme 3: Humanity could be greatly enfeebled by AI', *Experts Imagine the Impact of Artificial Intelligence by 2040* (Imagining the Digital Future Center, Elon University, Feb. 2024), pp. 20–28.

[41] Crawford, K., *The Altas of AI: Power, Politics and the Planetary Costs of Artificial Intelligence* (Yale University Press: New Haven, CT, 2021).

[42] Du, M. and Okola, C., 'Reimaging the future of data and AI labor in the Global South', Brookings, 7 Oct. 2025.

[43] Lemma, A., 'Critical minerals, critical moment: Africa's role in the AI revolution', ODI Global, 10 Feb. 2025.

particularly vulnerability. Studies have shown that malicious actors can exploit this overreliance by using low-resource languages (such as Zulu or Scots Gaelic, for which AI models have insufficient training data) to bypass or 'jailbreak' safety features and elicit harmful responses.[44]

A fourth scenario that has been a matter of great debate in AI and national security communities in recent years relates to the security implications of AI deployment and specifically to the question of whether datasets and algorithms that power the most advanced AI systems should be made openly available.[45] One of the concerns is that open access to such assets would level the international playing field in AI development, including in the military domain. Openness could empower military actors as well as malicious actors in ways that could be disruptive for international peace and security. Examples include making it easier to weaponize commercial drones or to enhance them with greater cyber capabilities. Alternatively, restricting openness could exacerbate tension between the *haves* and the *have-nots*, and lead some actors who feel strategically inferior in the AI domain to exercise power and threats through other disruptive means, including commercial measures or pursuing military capabilities that could be destabilizing for regional and global security.

Both perspectives are plausible, not least because the downstream consequences of both approaches are already being realized. Cases of misuse of open-source AI have already been reported.[46] There have also been cases where states respond to perceived technological asymmetries and technology access restrictions that are imposed on them by other states (e.g. via export control) with political and military measures.[47] The USA introduced export controls that restrict China's access to certain types of AI computer chips, to which China responded with, among others things, a resolution at the UN General Assembly critiquing multilateral export controls.[48]

The takeaway is that the development and diffusion of AI capabilities, regardless of how it is done, is poised to have geopolitical implications. It will affect how countries see their relative economic, military and political power, and also fuel reactions that could have negative consequences for international peace and security in the short or long term.[49]

Evaluating the risks

From a governance perspective, the manifold ways in which advances in AI pose risks to international peace and security are a challenge because they bring to the fore divergences on where to direct resources and what policy action should look like. Decision-makers need to know what risks deserve the most attention. This section

[44] Yong, Z-X., Menghini, C. and Bach, S. H., 'Low-resource languages jailbreak GPT-4', *arXiv*, 2310.02446v2, 27 Jan. 2024.

[45] Segun, S., 'The global security risks of open-source AI models', UNODA Responsible AI for Peace and Security Blog No. 6, 21 Feb. 2025.

[46] Ciancaglini, V. et al., *Malicious Uses and Abuses of Artificial Intelligence* (Trend Micro Research, UN Interregional Crime and Justice Research Institute, and Europol's European Cybercrime Centre, 2020); and Righetti, L. and Boulanin, V., 'Navigating the dual-use dilemma: Open access to research drives innovation, but how can we avoid unintended consequences?', *IEEE Spectrum*, 10 June 2025.

[47] Kashin, V. and Raska, M., 'Countering the US Third Offset Strategy: Russian perspectives, responses and challenges', RSIS Policy Report, 24 Jan. 2017.

[48] Shivakumar, S., Wessner, C. and Howell, T., 'The limit of chip export controls in meeting the China challenge', CSIS Commentary, 14 Apr. 2025.

[49] Schmid, S. et al., 'Arms race or innovation race? Geopolitical AI development', *Geopolitics*, vol. 30, no. 4 (2025); and Brockmann, K., Bromley, M. and Maletta, G., 'Implications of the UN resolutions on "international cooperation on peaceful uses": Balancing non-proliferation and economic development', SIPRI Topical Backgrounder, 11 Dec. 2024.

canvasses how experts view the spectrum of risks and where the dividing lines are in the debate.

Known versus speculative risks

Unsurprisingly, experts have diverging opinions on the risk landscape, and the level of agreement correlates closely to the speculative nature of the risk in question. The more direct and tangible the pathway to harm is, the less debate there is. The more the risk scenarios are built on predicting increases in AI capabilities and complex causal chains, the more experts disagree. It is useful in this context to divide the risk landscape into two categories—known/established risks versus speculative/hypothetical risks (see figure 2.1)—and discuss how experts approach each category in turn.

Misuse risks typically fall in the first category. The fact that some actors could misuse AI in harmful ways is well established and therefore beyond dispute. What is left for debate is less the likelihood of misuse, but rather how severe the impact of such misuses might be. To sample experts' views on this point, the authors of this report invited a group of internationally recognized experts from government, academia, civil society and industry to assess in a series of dialogues how AI would impact the current risk landscape in the political, cyber and physical domains. Experts were invited to rate their level of concern, on a scale of 0 to 5 (with 5 being most concerned), about misuse of AI in each domain, and then to indicate, for each domain, which of three descriptions best fits the qualitative impact of the latest advances in generative AI: (*a*) game-changing, as they introduced new threats in this domain (e.g. by enabling new varieties of attacks); (*b*) not game-changing, as they primarily expanded existing threats (e.g. by expanding the set of actors who are capable of carrying out an attack, increasing the speed at which an attack may be conducted, or increasing the number of vulnerable or plausible targets); or (*c*) not game-changing, as they affect the typical character of the threats (e.g. by making a certain type of attack more effective, more targeted or harder to attribute).[50]

Participants generally agreed that the political domain was the area where misuse of AI presented the most pressing risks (see figure 2.2). For the qualitative impact of AI in each domain, the answers varied between domains (see figure 2.3). The digital domain was the area where most experts perceived generative AI to have the most transformative impact, although perceptions were similar for the political domain, with the physical domain having the most variation of views.

The participants' subsequent discussion on the results of the survey illustrated the difficulty of discussing the impact of AI with experts from different disciplinary backgrounds. Familiarity bias tends to affect how experts evaluate the novelty and severity of certain misuse risks. AI experts, for instance, tend to stress the novel capabilities that an AI provides to threat actors, while experts from a specific domain tend to relativize the net effect of AI in that domain by comparing it to other or past technological developments. This difference of approach is also visible in the literature, for instance, in papers on how generative AI systems could enable the development and use of weapons of mass destruction. AI researchers are typically very worried that AI models could lower the knowledge barrier for designing, acquiring and using biological weapons.[51] Meanwhile, biosecurity experts are more likely to point out that several other important variables are at play when it comes to assessing biological weapons risks from the threat actors' intent, capabilities and financial resources (historical biowarfare programmes

[50] The question are based on the risk analysis framework develop by Brundage et al., *The Malicious Use of Artificial Intelligence: Forecast, Prevention and Mitigation* (note 2).

[51] Service, R. F., 'Could chatbots help devise the next pandemic virus?', *Science*, vol. 380, no. 6651 (2023).

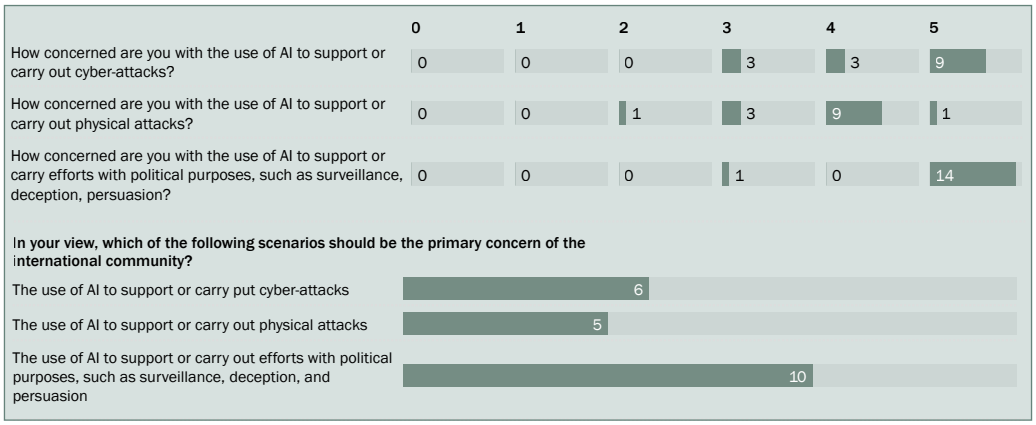

Figure 2.2. Levels of concern about misuses of artificial intelligence (AI), on a scale of 0 (lowest) to 5 (highest), as expressed by internationally recognized experts from government, academia, civil society and industry

have shown that biological weapons are neither easy nor cheap to produce).[52] For them, the fact that AI models make information more accessible and could generate steps for the development of dangerous pathogens is not, on its own, sufficient to change the biosecurity landscape. These divergences in risk assessment, however, do not eclipse the fact that there is a broad consensus among experts that the misuse of AI is a foreseeable tangible risk that requires an immediate response.

In this group discussion, but also in other dialogue activities that UNODA and SIPRI conducted, experts' views were much more divided when it came to accidental and structural risks. This is because these risk scenarios are, by design, more speculative. They require making a certain number of hypotheses about how AI capabilities might progress, but also how these capabilities will be deployed and the effects they will have. The loss of control scenario associated with superintelligent AI systems, for instance, is based on the premise that AI capabilities will continue to develop as quickly as they have lately and could further accelerate if AI reaches a point where it can autonomously self-improve.[53] Concerns around the structural impact of AGI are also based on the assumption that AI adoption will be very fast and wide-ranging: all sectors of the economy and society will want to leverage AI because the possibilities are so attractive.[54] Both premises remain highly debated in the AI community. Some AI experts (including Meta's chief AI scientist Yann Lecun) consider it flawed to extrapolate from the progress made in recent years, and that progress is likely to slow down because of (*a*) the lack of training data (there is only so much data to train on); (*b*) limitations around computational resources (data centres take time to build and also require a significant amount of water and energy to run); or (*c*) algorithmic limitations (LLM are unreliable and would need to be combined with new methods).[55] Others have also pointed out that even if AI capabilities continue to progress quickly, 'the transformative economic and societal impact will be slow (on the timescale of decades)', because there are numer-

[52] Lentzos, F., 'Artificial intelligence and biological weapons', T. Reinhold et al., 'Artificial intelligence, non-proliferation and disarmament: A compendium on the state of the art', EU Non-proliferation and Disarmament Consortium, Non-proliferation and Disarmament Paper No. 92, Jan. 2025, p. 22; and Mouton, C. A., Lucas, C. and Guest, E., 'The operational risks of AI in Large-scale biological attacks: Results of a red-team study', RAND Research Report No. RR-A2977-2, 25 Jan. 2024.

[53] Kwa, T. et al., 'Measuring AI ability to complete long tasks', METR Blog, 19 Mar. 2025.

[54] Kokotajlo, D. et al., *AI 2027*, AI Futures Project, 3 Apr. 2025.

[55] Lecun, Y., X, 1 June 2024 <https://x.com/ylecun/status/1796982509567180927>; and Marcus, G., 'The fever dream of imminent superintelligence is finally breaking', *New York Times*, 3 Sep., 2025.

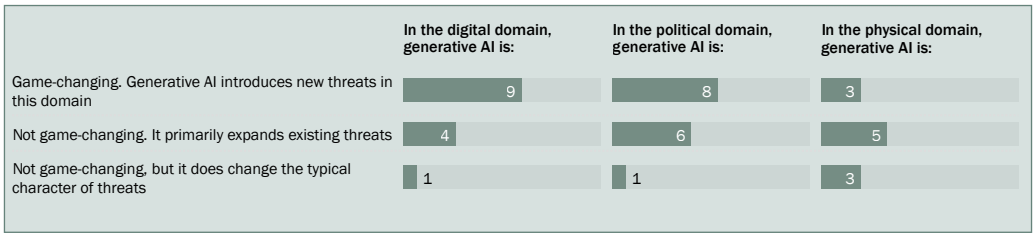

Figure 2.3. Qualitative impact of generative artificial intelligence (AI) threats in the political, digital and physical domains, as expressed by internationally recognized experts from government, academia, civil society and industry

ous obstacles to diffusion of the technology in society (from regulatory roadblocks to cultural and organizational factors).[56]

Experts' familiarity bias is also at play in the case of accidental and structural risks. AI experts tend to give more credit and importance to risk scenarios that are closest to their interests or the context in which they operate. Concerns around loss of control or large-scale replacement of human jobs by AGI are most often expressed by AI experts and AI organizations from the Global North. AI experts who focus on the impact of AI on the Global South—or who are from the Global South—often prioritize other types of concerns, not least how present-day development and use of AI systems are causing economic inequalities as well as societal and ecological disruption.[57]

The funding landscape has also impacted the type of risk scenario that receives expert attention. Over the past five years, there has been a lot of philanthropic interest in catastrophic risk associated with advanced AI systems. Funding organizations like Open Philanthropy, Founder's Pledge or Longview have disbursed a significant amount of funding (by some accounts, $110–130 million in 2024) towards research projects and activities that focus on speculative existential risks associated with advanced AI systems.[58] These organizations have been criticized for steering the research community's attention away from the more immediate harm from, for example, the societal impact of algorithmic bias.[59]

Risk prioritization: disagreement is no obstacle to action

In sum, experts disagree on the severity of misuse risk (not on its likelihood) and on both the likelihood and severity of structural and accidental risks. There are also debates on whether these risks are inherently novel. But do these differences of view matter? From a policy perspective, they are not fundamental obstacles for a few reasons.

First, disagreement among experts is an inherent feature of any technology impact assessment because of the experts' different backgrounds, levels of knowledge, values and interests. All the major technological breakthroughs of the 20th century led to intense and polarized debate. Expert debate always involves some form of tribalism, with specific camps, entrenched positions, and refusals to concede reasonable points because of particular worldviews or interests.[60] The same dynamics are at play today in the case of AI.

Second, behind the disagreements, there is also a lot of common ground. The experts consulted by the authors shared a common goal: to ensure that AI is developed and

[56] Narayanan, A. and Kapoor, S., 'AI as normal technology', Knight First Amendment Institute at Columbia University, 15 Apr. 2025.

[57] For instance, the Distributed AI Research Institute (DAIR) presents itself as an organization that explores 'real harms' and is critical of the focus on hypothetical risks scenarios. See DAIR, 'The real harms of AI systems', [n.d.].

[58] Quick Market Pitch, 'Who is funding AI safety research?', [n.d.].

[59] Gebru, T., 'Effective altruism is pushing a dangerous brand of "AI safety" ', *Wired*, 30 Nov. 2022.

[60] Russell (note 30), p. 159.

Box 2.3. Definition of responsible innovation

A frequently cited definition of responsible innovation is that of René von Schomberg: 'a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)'.[a] In the European context, the approach is often discussed under the label 'responsible research and innovation'. The distinction between research and innovation is intended to separate the issues related to fundamental science work (e.g. articles, conference papers and book chapters) from those that touch on applied work aimed at the development of commercial products and services.[b] For the sake of brevity, this report collates these two dimensions under the shorter and more commonly used label of 'responsible innovation'.

[a] von Schomberg, R., 'Definition of responsible innovation', René von Schomberg, [n.d.].
[b] Grunwald, A., *Technology Assessment in Practice and in Theory* (Routledge: New York, 2019), p. 20.

deployed responsibly, that is, in a way that mitigates potential harm and ensures socially desirable outcomes, including in the context of international peace and security.

Third, risk management is not a zero-sum game. As pointed out by Anca Dragan, director of AI Safety and Alignment at Google DeepMind, in a presentation at the inaugural conference of the International Association for Safe & Ethical AI, known and more speculative risks can and need to be addressed in parallel.[61] Addressing one type of risk does not need to be at the expense of another type of risk. The good news in that context is that, at some level, there is a common solution: implementing good practices in responsible innovation.

Addressing the risks: responsible innovation as a first step

Most, if not all, these risks can be prevented or mitigated through the greater use of responsible innovation practices within the AI community.

Responsible innovation refers to an anticipatory approach to technology governance that seeks to ensure that scientific and technical advances are steered towards beneficial outcomes and that negative impacts are identified and mitigated in advance (box 2.3). In practice, responsible innovation relies on principles, tools and processes that are intended to help all stakeholders that are involved in the processes of research and innovation to (*a*) identify risks and benefits associated with their work; (*b*) evaluate these risks and benefits in terms of likelihood and scale; and (*c*) use these considerations to limit or guide the design and development of new research, products and services (e.g. through functional requirements).[62] What makes responsible innovation so valuable is not only that it aims to prevent harm before the technology is deployed, but also that it is a technology- and risk-agnostic approach.[63] It can be applied to all sorts of AI systems and can elicit any type of risks and risk management response. Moreover, it is a methodology rather than a set of fixed principles or rules—there is no single recipe for responsible innovation—that can help practitioners navigate uncertainty around the impact of a given technology as well as the coverage of existing regulations.

Most of the AI risks outlined above, regardless of their nature, could be prevented or mitigated through the application of responsible tools and practices. Admittedly, responsible innovation is not a 'cure-all'—if users focus only on a single set of risks,

[61] Dragan, A., 'Navigating the path to AGI safely and responsibly', Speech, International Association for Safe & Ethical (IASEAI) Conference 2025, 6 Feb. 2025.
[62] van Oudheusden, M., 'Where are the politics in responsible innovation? European governance, technology assessments, and beyond', *Journal of Responsible Innovation*, vol. 1, no. 1 (2014).
[63] Boulanin, V., Brockmann, K. and Richards, L., *Responsible Artificial Intelligence Research and Innovation for International Peace and Security* (SIPRI: Stockholm, Nov. 2020).

responsible innovation practices will not automatically address other risks. However, if conducted properly with the right tools and right level of external input, responsible innovation practice can help AI practitioners and organizations identify potential downstream risks, including those outside of their immediate frame of reference, enabling them to adopt technical or procedural measures that can meaningfully reduce the likelihood or scale of a given risk's impact. What matters in that context is to ensure that individuals and organizations that develop and deploy AI apply the many available tools and processes more systematically and robustly.[64] This need is increasingly recognized and accepted both within the AI industry and the policy community. The following chapters describe how the AI community has so far made use of responsible innovation practice to address the risks that advances in AI may pose to international peace and security, and what policymakers could do (more or differently) to support such practices.

[64] In the context of AI, these tools and processes include 'IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being', *IEEE Std* 7010-2020 (2020); US Department of Commerce, National Institute of Standards and Technology (NIST), *AI Risk Management Framework* (NIST: Gaithersburg, MD, 2023); and European Commission, High-level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (European Commission: Brussels, 2019).

3. Responsible innovation practice in the AI community

Understanding the shift in awareness and engagement

The proposition that AI innovation can lead to harm and that such harm can be prevented and mitigated through responsible innovation has been widely recognized in the AI community for more than a decade. This proposition has even been crystallized into the concept of 'responsible AI', which is now a cornerstone of corporate and governmental policy (see box 3.1).

Major AI companies have typically articulated their commitment to responsible AI through the adoption of their own sets of principles and the establishment of specific teams, boards or committees, as well as investing in research around the safety, security, ethics and societal impact of AI.[65] Meanwhile, governments and international organizations have also strived to guide the AI community through the adoption of general principles (see chapter 4).[66] Responsible AI has emerged as a distinct topic for AI and interdisciplinary research.[67] How AI can be developed and deployed responsibly is now a common agenda item in academic conferences dedicated to AI.

For a long time, efforts on responsible AI had no direct or explicit connection to international peace and security. The type of harm typically in focus was social harm, such as harms stemming from algorithmic discrimination or labour force displacement. This was, in part, because in the eyes of many actors, concerns around AI's impact on international peace and security were exclusively to do with military AI. The civilian AI community, whose norms largely stem from computer science, was less familiar with dual-use issues and did not have a strong history of considering international peace and security risks outside of products developed for direct military contracts.

However, the situation has changed drastically over the past five years, due partly to a conjunction of events which contributed to bringing peace and security–related risks into the spotlight. First, the rapid democratization of generative AI tools led to a wave of studies looking at the harms that could come from possible use and misuse of generative AI systems.[68] Prominent figures in the AI community also issued or endorsed statements and open letters that called for greater attention to global risks associated with the rapid development and deployment of advanced AI systems.[69] Second, 2024 was a major election year with important elections around the world. Many analysts feared that AI would be misused to destabilize democratic processes.[70] While AI did not seem

[65] See e.g. Google AI, 'Our AI principles', [n.d.]; Microsoft, *The Microsoft Responsible AI Standard*, v2, June 2022; and IBM Responsible Technology Board, 'Reflecting on IBM's AI Ethics Board: Insights from the past 5 years for the future', IBM, 2024.

[66] See e.g. UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (UNESCO: Paris, 2021); NIST (note 64); and Organisation for Economic Co-operation and Development (OECD), 'OECD AI Principles overview', May 2024.

[67] See e.g. Dignum, V., *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (Springer: Cham, 2019); and Voeneky, S. et al. (eds), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (Cambridge University Press: Cambridge, 2022).

[68] See e.g. Marchal et al. (note 13); Shevlane, T. et al., 'Model evaluation for extreme risks', Google DeepMind, 25 May 2023; Weidinger et al, *Sociotechnical Safety Evaluation of Generative AI Systems* (note 14; Seger, E., et al., 'Open-sourcing highly capable, foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives', Centre for the Governance of AI, 2023; Goldstein et al. (note 4); Avin, S. et al., 'Filling gaps in trustworthy development of AI', *Science*, vol. 374, no 6573 (2021); and Anderljung, M., Hazell, J. and von Knebel, M., 'Protecting society from AI misuse: When are restrictions on capabilities warranted?', *AI & Society*, vol. 40 (2025).

[69] Future of Life (note 1); and Milmo (note 1.

[70] Kapoor, S. and Narayanan, A., 'Is AI-generated disinformation a threat to democracy?', AI as Normal Technology Blog, 19 June 2023.

Box 3.1. Responsible artificial intelligence

The concept of responsible artificial intelligence (AI) does not have a single universally applied definition. However, it is typically used to recognize that AI should be developed and deployed according to several high-level principles, namely by ensuring that AI systems are:

- *Safe*—AI systems are reliable, that is, they perform as intended, and are designed in a way that mitigates the possibility of, and consequences of, system failure(s).
- *Secure*—AI systems are designed in a way that minimizes their vulnerability to adversarial attacks and more broadly safeguards them from being misused for harmful purposes.
- *Legal*—AI systems' design and intended use allow the user to comply with existing laws and regulations. For instance, the systems do not, by design, breach data privacy laws or aim for uses that are prohibited, and do not preclude the user from complying with their legal obligations.
- *Ethical*—AI systems' goals and behaviours are aligned with human values and ethical standards. Some of these ethical standards have been formalized at the national or international level. The most prominent example is the adoption of UNESCO recommendations on the ethics of AI in 2021.[a]
- *Trustworthy*— AI systems' outputs are accurate and predictable.[b]
- *Socially desirable and sustainable*—AI systems' goals and behaviour fulfil societal needs and do so in a way that is not excessively at the expense of finite resources that are critical for society. In the context of international discussions on AI governance, the United Nations' 17 sustainable development goals can be considered an important reference point when determining what makes AI socially desirable and sustainable.[c]

[a] UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (UNESCO: Paris, 2021).

[b] European Commission, High-level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (European Commission: Brussels, 2019).

[c] United Nations, Department of Economic and Social Affairs, 'The 17 goals', [n.d.].

Source: Boulanin, V., Ovink, C. and Palayer, J., 'Handbook on responsible innovation in AI for international peace and security', UNODA Occasional Paper No. 45, July 2025.

to have a dramatic effect on the outcome of elections, the limited effect was, among other reasons, because the industry and policy community took preventive measures to mitigate the scale and impact of AI-enabled election interference and influence.[71] Third, the geopolitical landscape changed dramatically between 2020 and 2025. The outbreak of the wars in Ukraine and in Gaza, and the change of political leadership in the USA, affected how sections of the AI industry viewed the possibility that their work could have military end-uses or be misused for harmful ends. Civilian companies at the forefront of AI development, such as OpenAI, Google and Meta, are no longer reluctant to deal with military contracts and work on issues related to international peace and security.[72]

The nexus between AI and international peace and security now seems more broadly recognized within the AI industry. Most major AI companies have now published an 'AI Safety Framework' or equivalent official document (see box 3.2) that presents the types of risks associated with the development and deployment of their models and how they are responding to those at the company level. All these documents rank the risk of misuse (whether for influence operations, cyberattacks or weapon development) as

[71] For an overview of the other reasons see Simon, F. and Altay, S., 'Don't panic (yet): Assessing the evidence and discourse around generative AI and elections', Knight First Amendment Institute at Columbia University, 7 July 2025.

[72] Pascual, M. G., 'Big Tech enters the war business: How Silicon Valley is becoming militarized', *El Pais*, 21 July 2025.

Box 3.2. Safety frameworks published by major artificial intelligence companies

- Amazon's Frontier Model Safety Framework (2025) <https://www.amazon.science/publications/amazons-frontier-model-safety-framework>
- Anthropic's Responsible Scaling Policy (2023) <https://www.anthropic.com/news/anthropics-responsible-scaling-policy>
- Cohere's Secure AI Frontier Model Framework (2025) <https://cohere.com/security/the-cohere-secure-ai-frontier-model-framework-february-2025.pdf>
- G42's Frontier AI Safety Framework (2025) <https://www.g42.ai/resources/publications/g42-frontier-ai-safety-framework>
- Google DeepMind's Frontier Safety Framework (2024) <https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/>
- Magic's AGI Readiness Policy (2024) <https://magic.dev/agi-readiness-policy>
- Meta's Frontier AI Framework (2025) <https://ai.meta.com/static-resource/meta-frontier-ai-framework/>
- Microsoft's Frontier Governance Framework (2025) <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf>
- Naver's AI Safety Framework (2024) <https://clova.ai/en/tech-blog/en-navers-ai-safety-framework-asf>
- Nvidia's Frontier AI Risk Assessment (2025) <https://images.nvidia.com/content/pdf/NVIDIA-Frontier-AI-Risk-Assessment.pdf>
- OpenAI's Preparedness Framework (2025) <https://openai.com/index/updating-our-preparedness-framework/>
- xAI's Risk Management Framework (2025) <https://data.x.ai/2025-08-20-xai-risk-management-framework.pdf>

Note: For a systematic comparison see METR, 'Common elements of frontier AI safety policies', 26 Mar. 2025.

a priority. The companies' commitment to address these risks was also notable in the statements they made in recent policy events dedicated to AI safety, not least the AI safety summits in the United Kingdom (2023), Seoul (2024) and Paris (2025). Admittedly, the communication from major AI companies is not necessarily representative of the entire AI industry nor the wider AI community. Anecdotal evidence that the authors of this report collected in the context of an international workshop series with doctoral and masters students in STEM disciplines, and interviews with representatives from small and medium enterprises (SMEs), indicates that the level of understanding of how advances in AI can present risks to peace and security remains superficial and uneven. AI students and SMEs often lack the vocabulary and useful points of reference to understand how their work connects to international peace and security. Limited engagement and superficial understanding can in part be attributed to the fact that such issues are rarely taught in STEM education, and that SMEs typically do not have in-house capability to explore and articulate policies on how they perceive and address the risks associated with the products and services they sell.

AI community efforts to identify, evaluate and address risks

As alluded to in the previous section, it is challenging to capture the diversity of the AI community in a single picture—especially when many individual practitioners have multiple roles spanning academia, civil society and industry—but it is useful for these purposes to get a snapshot of current efforts. This section focuses on measures taken by major AI companies and industry organizations dealing with the most cutting-edge research and products, along with several observations on efforts made in the academic

sector. These companies, industry organizations and universities are primarily in the Global North, given the current centralization of AI development.

Over the past three years, the largest actors in the AI industry (such as Google, Microsoft, OpenAI and Anthropic) have been increasingly open about the types of measures they have put in place to identify, evaluate and address risks associated with their models and their products more generally. These measures can be grouped in three categories: (*a*) risk evaluation measures; (*b*) risk prevention and risk mitigation measures; and (*c*) information sharing and collaborative measures with the wider AI industry, civil society and government.

Risk evaluation measures

Major AI companies have put significant effort into developing risk evaluation methodologies to identify and evaluate the risks associated with their models, from known misuse risks to more speculative risks. Typically, they rely on two kinds of methodologies: benchmarking and red-teaming.

Benchmarking is like a standardized academic test for AI models. It is a question-based evaluation process which aims to measure and score a model's capabilities in a specific area. This is the method that is commonly used to compare AI performance with human performance and to assess whether models are getting better at a certain subject or task over time.

In the context of risks to international peace and security, companies have been using benchmarking to evaluate their models' knowledge and capabilities in the context of a specific misuse scenario: what can the model do to assist the development and conduct of a large-scale influence operation or cyberattack, or to support the production and use of chemical, biological, radiological and nuclear (CBRN) weapons. Companies typically develop these benchmarks in-house or in collaboration with independent researchers or expert organizations or, in cases where classified knowledge is required, governments.[73] The process typically entails asking domain experts (e.g. cybersecurity, biosecurity, nuclear physics) to generate a large set of question-and-answer pairs to test the model's knowledge on topics spanning different levels, from undergraduate/postgraduate information to specialized open-source knowledge known to domain experts but not widely publicized. Some independent actors have also started developing their own benchmarks that companies, governmental and independent evaluation organizations can use to assess the models on sensitive topics.[74]

Benchmarking as an approach has both benefits and limitations. A strength is that the process can be automated (the answers provided by the model can be analysed by an AI classifier), and it also provides a means to measure and compare AI models' harmful capabilities. The downside of that approach is that its effectiveness is very much dependent on the quality and comprehensiveness of the questions provided. Moreover, it does not really permit testing the depth of the knowledge of the model, and when not sufficiently rigorous, it is vulnerable to misleading results (e.g. if benchmark questions and answers are already included in the dataset).

These limitations are the reasons AI companies also use *red-teaming* as an additional evaluation technique. The concept is borrowed from the field of cybersecurity, where it usually refers to a process involving conducting realistic attacks on systems to test for vulnerabilities and to understand likely adversary capabilities and goals. In the AI context, red-teaming refers to an interactive process where experts (or an AI) probe

[73] Sabin, S., 'Exclusive: Anthropic, feds test whether AI will share sensitive nuke info', Axios, 14 Nov. 2024.

[74] See e.g. Noever, D. and McKee, F., 'Forbidden science: Dual-use AI challenge benchmark and scientific refusal tests', *arXiv*, 2502.06867, 8 Feb. 2025; and Rein, D. et al., 'GPQA: A graduate-level Google-proof Q&A benchmark', *arXiv*, 2311.12022, 20 Nov. 2023.

the AI model in different ways to identify harmful capabilities or outputs, but also to evaluate how vulnerable the system is to adversarial attack (jailbreaking attacks).[75] The companies report that they find red-teaming a particularly valuable methodology because it is, by design, a dynamic, iterative process that can not only help prioritize risk mitigation strategies, but also help determine the effectiveness of these strategies, because it can be repeated over time.[76] Companies now routinely use red-teaming to assess how malicious actors might misuse their AI model and to evaluate the effectiveness of the model's safeguards. The way they do this varies, however. Microsoft, for instance, does it in-house using internal experts, while OpenAI and Anthropic have worked with external subject-matter experts and organizations.[77] Some companies have also experimented with LLM to automate red-teaming. Whether it is wise to rely on an AI system to probe harmful content remains a debated issue in the literature. Proponents of AI solutions point to the fact that AI systems may have limitations but generate, on average, better results—in the sense of more consistent and faster results—than human red-teaming efforts.[78]

Red-teaming, like benchmarking, has inherent limitations.[79] These include:

- *Resources*. Red-teaming is resource-intensive. It takes time and requires a lot of human resources.
- *Human performance*. The quality of the red-teaming process is very much dependent on the quality of the human expertise that is mobilized. One of the lessons that OpenAI learned from red-teaming GPT4 is that human experts showcased varying levels of creativity, motivation and capability during the exercise.
- *Lack of standards*. The research community lacks shared norms, best practices and technical standards for how to safely and effectively red-team AI systems. This is particularly true for CBRN weapons risks.[80]
- *Security clearance versus information sharing*. Some actors may need to receive some form of security clearance. Others who do have security clearance from their employer may not be allowed to draw on the information they have.
- *Legal safeguards*. The rules are not clear for people and organizations that seek to conduct independent red-teaming of AI models. AI companies' terms of service and use typically disincentivize or make it practically impossible for independent 'white hats' to red-team deployed models.
- *Test versus reality*. One of the lessons that OpenAI learned from deploying GPT2 and GPT3 was that there was a mismatch between expected misuses and forms of misuse encountered in the wild. The original red-teaming focus was on the generation of malware and misleading political content. They had not expected a spam proposal for a dubious medical product and

[75] Frontier Model Forum, 'What is red teaming?', Issue brief, 24 Oct. 2023.
[76] See e.g. Anthropic, *Responsible Scaling Policy*, v2.2, 14 May 2025.
[77] Frontier Model Forum, 'What is red teaming?' (note 75).
[78] Perez, E. et al., 'Red teaming language models with language models', *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2022); and Ganguli, D. et al., 'Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned', Dataset, [n.d.].
[79] Brundage, M. et al., 'Lessons learned on language model safety and misuse', OpenAI, 3 Mar. 2022.
[80] Brundage et al., 'Lessons learned on language model safety and misuse' (note 79).

roleplaying of racist fantasies.[81] The takeaways here are that it is difficult for a red-team exercise to foresee all possible misuses.

Risk prevention and mitigation measures

AI companies typically use the insights from the risk evaluation process to then deploy various types of risk prevention and risk mitigation measures. These measures fall within one of two categories according to whether they focus on the design or the deployment of the AI.

Design-related interventions can take many forms but are technical in nature and aim to increase the safety, security or social acceptability of the systems during the development stage. Companies rely extensively on these types of measures to prevent possible misuse or reduce the harm associated with potential misuses.[82] In the case of a general-purpose chatbot, measures include using an input filter (itself generally an AI tool that needs to be trained) to ensure the dataset on which the model is trained does not include harmful or sensitive information, as well as an output filter to ensure that the model does not share harmful information. In the context of autonomous navigation in drones or vehicles, measures include geo-fencing to prevent travel to protected or sensitive areas.

Deployment-related interventions, in contrast, aim to either affect the willingness and ability of people to potentially misuse a deployed system or to limit the exposure of users to potential harm. Companies have employed a panoply of measures ranging from asking users to identify themselves with email or governmental ID; restricting the user base to only certain types of certified or authorized users; restricting access based on geo-location; limiting the number or the size of the requests that a user can make to the system; monitoring usage to detect possible misuse; and applying a strict policy on 'acceptable use' (e.g. by explicitly prohibiting certain usages) that includes revoking product licences or denying software upgrades to users that break the policy.

Companies that authors engaged with in the context of the project (including Microsoft, OpenAI, Google DeepMind, Amazon, Anthropic and Boston Dynamics) reported that, thanks to this wide spectrum of measures, they have been able to detect or quickly stop many forms of misuse. At the same time, they also recognized that it was difficult to foresee all the ways malicious actors might try to misuse their product or to bypass technical safeguards and usage restrictions. Calibrating technical safeguards and usage restrictions in a way that does not disproportionally affect the performance of the system is also difficult. For instance, in the case of a general-purpose AI chatbot, it is common for the output filters to block benign uses because they are too sensitive to certain words. Moreover, the filters can also introduce some latency in the production of the output because they often involve relying on another AI system running alongside the main model.[83] Misuse prevention and mitigation is, in that respect, also an engineering challenge. It requires experimenting with different methods and adapting them over time. In some cases, it may also require new methods. Boston Dynamics, for instance, developed its own methods to detect when its robot dogs are weaponized or used to fire a weapon, or are even in an environment with gunfire.[84]

[81] Brundage et al., 'Lessons learned on language model safety and misuse' (note 79).
[82] Anderljung, Hazell and von Knebel (note 68).
[83] Anderljung, Hazell and von Knebel (note 68).
[84] 'Episode 4: Boston Dynamics: How to deal with possible misuse of general purpose robots?', UNODA/SIPRI Responsible AI for Peace and Security Podcast series, 17 Apr. 2024.

Information sharing and collaboration

There seems to be broad agreement among AI companies and AI experts that the field of AI risk prevention and risk mitigation is still an emerging science.[85] Methods to identify possible misuses and dangerous capabilities in advanced AI models are still very new and can therefore be improved. In this vein, some major AI companies have taken steps to foster greater information sharing and collaboration with each other, and to some extent with independent organizations and governments.

One such step was the creation of the Frontier Model Forum in July 2023 on the joint initiative of Anthropic, Google, Microsoft and OpenAI. The forum is a membership organization for companies that have developed advanced general-purpose AI models (so-called frontier models). It was created to facilitate collaboration between member companies on safety research and establishment of best practices and shared standards for the responsible development and deployment of these systems.[86] The forum was also created to support information sharing with government, academia, civil society and the wider industry. To that end, the Frontier Model Forum has already published several briefs that allow public insight into what member companies have been doing to address the risks associated with their models. It has been an important source of information for this report. Another noteworthy development was that many leading AI companies—including from outside the USA—committed at the 2024 AI Safety Summit in Seoul to develop and publish safety frameworks that present the measures they deploy at the company level to mitigate risks in the development of advanced models.[87]

However, the information each company provides via the Frontier Model Forum, their safety framework or website remains very general. Technical details about the methodologies that the company uses—for instance, evaluating the risk that their model could assist with the development of CBRN weapons—are limited. Independent experts have therefore called for more transparency on what companies test, how they conduct those tests, and how they use the results to inform decisions.[88]

The question of whether companies should be more transparent about their methods is sensitive. In a closed-door dialogue that the authors organized with representatives from major companies, several of them pointed out that their companies had to account for potential information hazards.[89] Disclosing too much detail about how evaluations are conducted or how safeguards work could provide malicious actors with critical information to bypass these measures. For those reasons, the same representatives explained that their companies preferred sharing information only with other companies and relevant governmental agencies, such as the national AI safety and security institutes created following commitments made by governments at the 2023 and 2024 AI safety summits. The representatives viewed the Frontier Model Forum as the most useful venue for engagement and information sharing among the companies. They also pointed out that the national AI institutes could provide formal reporting channels and serve as hubs for know-how on certain domains, such as those pertaining to national or international security.

[85] Koessler, L. and Schuett, J., 'Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries', *arXiv*, 2307.08823, 17 July 2023.

[86] Frontier Model Forum, 'About us', [n.d.].

[87] British Department for Science, Innovation and Technology, 'Frontier AI safety commitments, AI Seoul Summit 2024', Policy paper, 21 May 2024.

[88] See e.g. McCaslin, T. et al, 'STREAM (ChemBio): A standard for transparently reporting evaluations in AI model reports', *arXiv*, 2508.09853, 3 Sep. 2025.

[89] Bostrom's definition of information hazard: 'a risk that arises from the dissemination or the potential dissemination of (true) information that may cause harm or enable some agent to cause harm'. Bostrom, N., 'Information hazards: A typology of potential harms from knowledge', *Review of Contemporary Philosophy*, vol. 10 (2011).

Academia: responsible research and publication

Responsible innovation practice in academia has been more difficult to map out for practical reasons of scale—the full spectrum of academic efforts on responsible innovation includes both the practices employed and the education provided in support of such practices. Doing so internationally would hardly be feasible and beyond the scope of this report. However, several high-level observations can be made based on the in-person workshops and dialogue activities that the authors conducted with professors and STEM students from around the world between 2023 and 2025.

The first is that academia faces a different set of problems than industry when it comes to identifying and mitigating the risks that AI can present to international peace and security. Most AI laboratories within universities tend to focus on fundamental research rather than the development of commercial products. The ways in which their work could have negative downstream consequences for peace and security are therefore less straightforward. Often, much of the knowledge that scholars produce is theoretical and therefore not directly misuse-ready. As many of the doctoral students at the workshops organized by the authors of this report observed, it requires quite some imagination for researchers to think of how their work could find harmful downstream uses.

The second observation is that applied research projects have a clearer misuse potential. An example is a new algorithm for controlling a drone swarm. Publishing that algorithm on an open-source software repository (like GitHub) would mean making it available to militaries and malicious actors. According to the professors in robotics and AI that the authors engaged with, the academic ecosystem provides too few means and incentives for researchers to think systematically of such a possibility.[90] Large universities often have a review board that can help scholars assess whether their work could be misused, be subject to export control, or pose difficult ethical questions. However, these review boards typically need to be actively consulted, and it is reportedly not uncommon that scholars, especially junior ones, ignore the possibility, and sometimes the duty, to consult with them before making their work publicly available.

The third observation is that STEM education typically provides little to no means for scholars to develop the knowledge and skills needed to assess the possible negative downstream consequences of their work, including potential misuse.[91] In other words, STEM education generally does not cover responsible innovation practices. Some universities (such as New York University, Delft University of Technology, Umeå University and Tallinn University of Technology) have made some efforts to mainstream consideration for ethics and responsible innovation, but these remain exceptions rather than the norm.[92] The extent to which academics have the flexibility to provide such education also varies significantly around the world, due to factors ranging from central determination of curriculums by government entities to pressures on the course load. A related problem is that efforts that scholars might put into assessing and mitigating potential misuse risks are not rewarded by the academic promotion systems. The 'publish or perish' pressure requires emerging scholars to prioritize work and considerations that are valued in their respective fields, and disincentivizes them to exercise extra caution in the disclosure of the work.[93] For similar reasons, students

[90] Boulanin, V., et al., 'AI missteps could unravel global peace and security—to mitigate risks, developers need more training', *IEEE Spectrum*, 24 July 2024.

[91] Dignum, V., 'The role and challenges of education for responsible AI', *London Review of Education*, vol. 19, no. 1 (2021).

[92] Figaredo, D. D. and Stoyanovich, J., 'Responsible AI literacy: A stakeholder-first approach', *Big Data & Society*., vol. 10, no. 2 (2024).

[93] Righetti and Boulanin (note 46).

and emerging scholars in STEM are rarely incentivized to engage in interdisciplinary dialogue (e.g. with peers from social sciences and law faculties) despite this being a useful way for them to not only identify potential risks but also to make their work more desirable and beneficial from a societal standpoint.

The idea that researchers and engineers should think more proactively about the potential societal impact of their work seems, all in all, rather widely accepted in academia. The real divergence of opinion revolves around how and to what extent such considerations should guide decisions around the publication of their work. In this context, over the past five years, several prominent AI figures and organizations have debated how individual researchers and organizations could engage in 'responsible publication'.[94] An important element of guidance that emerged in the debate was that there are ways to balance the need for openness (verifiability and reproducibility of research) with the need to prevent potential malicious downstream use of the research.[95] However, it is important not to reduce this debate to a dichotomy between open versus closed. There is a large menu of options that can be explored to find the appropriate balance so that academics do not have to choose between making their work fully accessible or fully closed. Here are some of the measures that academics can take to reduce risks associated with the research:

- *Leverage insights from peer review*: The peer review process provides an opportunity to gather insights from peers on the possible negative downstream consequences of the publication or diffusion of the work.
- *Consult with the relevant ethics entity within the university*: Universities often have a board or staff dedicated to thinking about ethical, legal and security implications of the work done at the university.
- *Curate the information*: Details that could be problematic could be identified and omitted from the publication version of the research. The authors of the frequently cited paper on the dual-use risk associated with AI-powered drug discovery used that approach.[96] They only presented the general logic of the experiment, rather than the full details. Similarly, implementing techniques like differential privacy can protect individuals' data while still allowing statistical analysis.
- *Limit functionality*: Applied researchers and developers could release a version of the work with reduced capabilities to limit potential misuse.
- *Conduct due diligence around collaborative research*: Researchers can ensure that research collaborators are trusted partners who will not share sensitive aspects of the work.

What challenges remain

There is clear evidence that the level of awareness and engagement of the civilian AI community on issues at the nexus between AI and international peace and security has increased. The section above did not provide a comprehensive picture, but the actions taken by major players in the AI industry over the past five years show progress in the right direction. Industry actors are taking tangible steps to prevent and mitigate risks of misuse and other harmful scenarios that could be associated with their products,

[94] Partnership on AI, 'Managing the risks of AI research: Six recommendations for responsible publication', White paper, 6 May 2021.
[95] Seger et al. (note 68).
[96] Urbina et al. (note 18).

some of which have a direct connection to international peace and security. Challenges remain, however.

First, as pointed out above, the methods that companies have applied to evaluate and deal with the different risks are still in their infancy. There is room not only for improvement of the various methods but also for greater harmonization across the industry on the criteria that are essential to their effective implementation. The Frontier Model Forum has pointed to the difficulty of formulating the critical thresholds that determine the acceptability of certain risks—for example, at what points does a general-purpose AI 'significantly' improve the capability of a malicious actor to develop biological weapons?[97] Moreover, the field of AI is advancing quickly. The shift to agentic AI—AI systems capable of acting in the digital or physical environment without direct human guidance over a long-time horizon—for instance, demands the development of specific risk evaluation procedures as well as technical safeguards.[98]

Second, recent years have also shown that a company's commitment to developing and deploying AI responsibly can change rapidly because of market pressures or a change in the political environment. The release of ChatGPT by OpenAI in late 2023 reportedly led several companies that were working on similar models, not least Google and Anthropic, to ship their product more quickly.[99] Many people who worked on AI safety and risk evaluation at OpenAI left the company after the launch of ChatGPT due to disagreement on the level of care (or lack thereof) that the company applied regarding the conduct of risk evaluation and the implementation of technical safeguards.[100] The change of administration in the USA in 2025 also politicized the topic of responsible development and deployment of AI, with campaign materials dubbing the previous administration's Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of AI as 'impos[ing] Radical Leftwing ideas on the development' of AI.[101] At the AI Action Summit in Paris, the new US administration, through the vice president, argued that some safety safeguards that companies had applied, in part in response to the EU Digital Services Act, were not only limiting free speech but were also partisan and damaging innovation.[102] Around the same time, some companies (e.g. Meta, YouTube) announced that they would (further) loosen up some of the restrictions they had placed on their AI systems.[103] Such changes in company policy have raised concerns that companies could engage in a 'race to the bottom' on AI safety and security, making general-purpose models easier to misuse and less safe to use. In contrast, the EU has been pulling in a different direction. The August 2024 entry into force of the Artificial Intelligence Act (EU AI Act) is placing novel requirements on

[97] Frontier Model Forum, 'Thresholds for frontier AI safety frameworks', Issue brief, 7 Feb. 2025; and Frontier Model Forum, 'Frontier AI biosafety thresholds', Issue brief, 12 May 2025.

[98] Boulanin, V., Blanchard, A. and Lopez Da Silva, D., 'Before it's too late: Why a world of interacting AI agents demands new safeguards', SIPRI essay, 1 Oct. 2025.

[99] Remmelt, 'Anthropic's leading researchers acted as moderate accelerationists', *LessWrong*, 2 Sep. 2025; Grant, N. and Weise, K., 'In AI race, Microsoft and Google choose speed over caution', *New York Times*, 7 Apr. 2023; and Weise, K. et al., 'Inside the AI arms race that changed Silicon Valley forever', *New York Times*, 5 Dec. 2023.

[100] Hao, K., *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI* (Penguin Press: New York, 2025).

[101] Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Executive Order 14110 of 30 Oct. 2023, *Federal Register*, 88 FR 75191, 1 Nov. 2023; and Trump 2024 Presidential Campaign, '2024 GOP platform: Make America great again!', [n.d.], p. 9.

[102] Vance, J. D., 'Remarks by the Vice President at the Artificial Intelligence Action Summit in Paris, France', 11 Feb. 2025.

[103] Isaac, M. and Schleifer, T., 'Meta says it will end its fact-checking program on social media posts', *New York Times*, 7 Jan. 2025; Grant, N. and Mickle, T., 'YouTube loosens rules guiding moderations of videos', *New York Times*, 9 June 2025; and Horwitz, J., 'Meta's AI rules have let bots hold "sensual" chats with kids, offer false medical info', Reuters, 14 Aug. 2025.

companies when it comes to evaluating and mitigating potential risks associated with general-purpose models.[104]

Third, when it comes to engaging in responsible development and deployment of AI, the playing field is uneven. There is a big capacity gap between the large US-based AI companies and the rest of the industry. Large US corporations have the financial resources to recruit multidisciplinary teams that can work on safety, security, ethics and model evaluation. They can even afford to employ domain experts to assess the capabilities of their models in biology or chemistry. Meanwhile, SMEs, which represent the largest share of the AI industry in Europe, typically have little to no resources dedicated to developing and maintaining procedures for spotting and stopping potential misuses and other negative downstream impacts of their products. This point came across clearly in an online roundtable that the authors held with representatives from SMEs, who expressed the difficulty of implementing comprehensive risk assessment processes due to their limited resources. They advocated for formalized risk management frameworks tailored to SMEs, which could take the form of checklists, self-assessments, and basic end-user screening protocols which would be accessible and practical for organizations without dedicated compliance teams.

A final and related challenge is that STEM education in most universities does not adequately prepare future AI practitioners—whether researchers, engineers or managers—for responsible innovation practice. Responsible AI requires a spectrum of capabilities that are typically not covered in STEM education. Courses on the societal impact of technology and responsible innovation, as well as specific training on AI ethics and governance, could help ensure that the AI practitioners of tomorrow can not only innovate more responsibly but also be more meaningful contributors and implementers of AI regulations.[105]

[104] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), *Official Journal of the European Union*, L series, 12 July 2024; and European Commission, The General-Purpose AI Code of Practice, 10 July 2025.

[105] Boulanin et al. (note 90).

4. How states and international organizations can support responsible innovation for international peace and security

The rapid pace of AI development, particularly for general-purpose models, has brought to the forefront a critical issue: responsible innovation is a collective action problem.[106] While it is in everyone's long-term interests for AI to be developed responsibly, individual actors—especially companies in a competitive market—have strong incentives to prioritize self-interests (e.g. be first to deploy a product or publish a research paper on a critical topic) over safety. As AI safety researchers from OpenAI noted years ago, competitive pressures can lead to underinvestment in safety and security measures, potentially leading to a dangerous race to the bottom. These competing incentives mean that industry self-governance alone is not sufficient to ensure that risks are identified and addressed promptly. Internationally coordinated governmental interventions are therefore necessary to establish minimum standards across the industry, ensuring that AI is developed and deployed responsibly. This chapter argues that effective international governance does not impede innovation; rather, it is a necessary foundation for responsible innovation that can help to prevent the misuse of AI and mitigate other risks that AI may present to international peace and security. Building on multi-stakeholder dialogues, industry roundtables and interviews conducted by the authors, this chapter explores how the AI community approaches the need for regulation and their relationship with regulators—that is, states and international organizations—looking at areas of agreement and disagreement. It also analyses nascent initiatives by states and international organizations, identifies remaining governance gaps, and proposes concrete steps for strengthening responsible innovation in AI for international peace and security.

How the AI community approaches the need for governance

The need for governmental intervention is now widely recognized, even among industry actors. Companies like Anthropic and Google DeepMind have publicly supported the creation of national AI safety institutes. There seems to be a strong consensus on the need for states and international organizations to support technical work in AI safety. This includes not only funding foundational research into areas like model interpretability and robustness, but also supporting the improvement of risk evaluation methods for dangerous capabilities. Scholars and experts from organizations like the Oxford Martin AI Governance Initiative, the Centre for the Governance of AI and the Partnership on AI have consistently advocated for a public research agenda to complement private sector efforts, arguing that safety is a global public good that commercial incentives alone will not fully address.[107]

Disagreements arise over the specifics of governance. A key point of contention is the balance between industry self-regulation and governmental intervention. Large AI companies typically argue they are best positioned to develop safety protocols due to their technical expertise. They fear that heavy-handed government regulation could

[106] Askell, A., Brundage, M. and Hadfield, G., 'The role of cooperation in responsible AI development', *arXiv*, 1907.04534, 10 July 2019.

[107] Blomquist, K. et al., 'Examining AI safety as a global public good: Implications, challenges, and research priorities', Carnegie Endowment for International Peace, Concordia AI and Oxford Martin AI Governance Initiative working paper, 11 Mar. 2025; Anderljung, M., et al., 'Frontier AI regulation: Managing emerging risks to public safety', *arXiv*, 2307.03718, 7 Nov. 2023.

'stifle innovation' and give an advantage to actors in less regulated jurisdictions.[108] Others, especially in academia and civil society, argue that governments must establish baseline requirements to ensure a level playing field, preventing a race to deploy without adequate safety checks.[109] While initiatives like the Frontier Model Forum demonstrate the industry's commitment to self-governance, the pace at which major AI companies have deployed general-purpose models validates the concern that self-governance efforts may not be sufficient to ensure all risks are identified and addressed in time.[110]

Another debated issue pertains to the locus of regulation—whether regulation efforts should be primarily national, regional or international. The world's largest AI companies have indicated a strong preference for light-touch governmental intervention at the national level, rather than through international organizations like the UN.[111] While they acknowledge that international coordination ensures regulatory consistency and prevents governance loopholes that bad actors could exploit, they have questioned the ability of a UN-led process or agency to be sufficiently flexible and responsive to address the challenges. According to scholars and representatives from civil society, companies' reluctance to see an international governance framework flows in part from economic considerations and the fear that international regulation could limit their room to manoeuvre or require them to comply with a costly compliance mechanism.[112] Across all sectors, AI practitioners from countries that are not major sites of AI development or deployment seem more positive to the emergence of an international governance regime, not least because many see it as the only way to correct some of the structural risks stemming from AI that disproportionately affect the Global South.[113]

A third and related area of disagreement is around whether governance intervention should take the form of legally binding measures, politically agreed norms or voluntary commitments. Again, industry actors favour voluntary pledges because they offer more flexibility, while actors from civil society and academia have been more vocal about the need for regulation that includes clear enforcement mechanisms. That divergence emerged clearly in the context of the adoption of the EU AI Act and whether companies should be legally mandated to evaluate how their model may be misused to support the development of weapons of mass destruction. After weeks of intense discussions, in which AI companies lobbied against legal requirements, the member states of the EU agreed to develop a code of practice for general purpose AI, but on the condition that it

[108] De Vynck, G. and Tiku, N., 'AI execs used to beg for regulation. Not anymore', *Washington Post*, 8 May 2025; and Marcus, G., 'Two years ago today in AI history: The tale of an about-face in AI regulation', Marcus on AI Blog, 16 May 2025.

[109] Tugend, A., 'Experts on AI agree that it needs regulations. That's the easy part', *New York Times*, 6 Dec. 2023.

[110] Toner, H. and McCauley, T., 'AI firms mustn't govern themselves, say ex-members of OpenAI's board', *The Economist*, 26 May 2024.

[111] Views expressed in a series of closed-door virtual roundtables that the authors organized with representatives from from major AI companies during spring 2025. Views were expressed under Chatham House rules and are therefore not attributable. For OpenAI's and Google's statements to government see OpenAI, 'OpenAI's proposal for the US AI Action Plan', 13 Mar. 2025; Jain, V., 'Google's recommendation to regulate AI', JustAI, 7 Aug. 2024; and Kent, W., '7 principles for getting AI regulation right', Google Policy Statement, 26 June 2024.

[112] Views expressed in a series of virtual multistakeholder dialogues that the authors organized with representatives from academia, civil society and private sector over the course of 2024. Views were expressed under Chatham House rules and are therefore not attributable. See also Hine, E., 'Artificial intelligence laws in the US are feeling the weight of corporate lobbying', *Nature*, 18 Sep. 2024.

[113] The Group of 77 at the United Nations, Statement on behalf of the Group of 77 and China delivered by the Iraqi delegation during the intergovernmental process and consultations to identify the terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence, New York, 17 Jan. 2025.

would be a voluntary tool for providers of such models to demonstrate compliance with the EU AI Act.[114]

National and multilateral governance efforts

Major initiatives so far

Until recently, the risks that advances in civilian AI pose to international peace and security were largely falling into a governance gap. These issues were not properly addressed in multilateral discussions aimed at governing either development or use in the civilian or military domain. However, that is changing. The combination of dynamic technological progress and recognition of the collective action problem has pushed states to act. Over the past two years, there have been several national and multilateral policy initiatives that seek not only to address some of the risks that AI presents to peace and security, but also to make the AI community and AI industry consider these risks more systematically as they research, develop or deploy AI. The three most significant initiatives are:

1. *The EU AI Act*: This landmark piece of legislation categorizes AI systems by risk level. While it is a regional initiative, it has global implications, as it sets baseline requirements for organizations that develop and deploy AI, including risk assessment and misuse prevention. However, the compliance requirements that the Act places on general-purpose AI via the Code of Practice are not legally binding, but voluntary, which means there are no financial sanctions attached to cases of non-compliance. Reportedly, one of the reasons for not making risk prevention measures mandatory was that methods for risk evaluation, prevention and mitigation were still in their infancy, such that it would have been premature to mandate processes that were not yet fully proven or established.[115]
2. *The Group of Seven (G7) Code of Conduct*: This is a significant multilateral effort that sets out a voluntary code of conduct based on a series of voluntary guiding principles for advanced AI development, including risk management policies and the need for robust security controls (box 4.1).[116] What is remarkable about the G7 Code of Conduct is that it describes a set of practices to identify and mitigate risks across the AI development and deployment lifecycle. It can serve as a baseline for international best practice around responsible innovation in AI.
3. *AI Safety Summit series*: These summits are high-level meetings that bring together governments, leading AI companies and researchers to discuss AI safety risks and potential mitigation strategies.

The EU AI Act and the G7 Code of Conduct are significant because they moved beyond aspirational principles around responsible AI (such as those formulated in the UN Educational, Scientific and Cultural Organization (UNESCO) *Recommendation on Ethics of AI* in 2021) to set practical requirements for organizations that develop and

[114] Nezim, P., 'How US firms are weakening the EU AI code of practice', *Tech Policy.Press*, 30 June 2025; and 'Tech lobby group urges EU leaders to pause AI Act', Reuters, 26 June 2025.

[115] Gallese, C., 'The GPA Code of Practice, a long journey not over yet', *MediaLaws*, 8 Sep. 2025.

[116] Group of Seven (G7), Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, 2023; and European Commission, 'Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI system'.

Box 4.1. G7 Code of Conduct: recommendations to organizations that develop and deploy advanced artificial intelligence (AI) systems

1. Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate and mitigate risks across the AI lifecycle.
2. Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market.
3. Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increase accountability.
4. Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia.
5. Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach—including privacy policies, and mitigation measures.
6. Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards across the AI lifecycle.
7. Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible, such as watermarking or other techniques to enable users to identify AI-generated content.
8. Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.
9. Prioritize the development of advanced AI systems to address the world's greatest challenges, notably but not limited to the climate crisis, global health and education.
10. Advance the development and, where appropriate, adoption of international technical standards.
11. Implement appropriate data input measures and protections for personal data and intellectual property.

Sources: Group of Seven (G7), Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, 2023; and European Commission, 'Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI system'.

deploy AI.[117] They articulate expectations around responsible innovation practices, including aspects that pertain to specific misuse cases (like CBRN weapons) and accidental harmful outcomes. Another important notable element is that they were framed—at least from the perspective of the drafters—as an enabler of responsible innovation, not as obstacles. They were meant to create a predictable framework for safety and responsibility that enables companies to compete on an even playing field. The AI safety summit series is also remarkable in the sense that it dedicated—at least in the first two summits—significant policy attention to large-scale risk that could flow from AI misuse, technical failure or emergence of dangerous capabilities in advanced AI models. The series also provided a forum where both states and companies could make official comments on AI safety. Several states, not least the USA, the UK, South Korea, Japan and, more recently, France, announced at the summits that they would create a national institute dedicated to AI safety and security. Major AI companies also committed to developing an internal framework for the responsible development and deployment of AI models, in which they would articulate the protocol they follow to assess and mitigate large-scale risks associated with their models (see chapter 3 and box 3.2).

[117] UNESCO (note 66).

Enduring challenges

Despite these commendable efforts, national and multilateral conversations have been hampered by several persistent challenges at three levels: conceptual, institutional and geopolitical.

Conceptually, the recommendations of the EU Code of Practice and the G7 Code of Conduct for the responsible development and deployment of AI remain very general. They provide little concrete guidance as to how AI models should be evaluated for potential safety, security and societal risks, such as which risks should be prioritized, how to assess these risks technically, and what expertise and resources are needed in the process. As a result, AI practitioners and companies need to do a lot of interpretative work to put these principles into practice.

Institutionally, the issue of how advances in civilian AI can impact peace and security still falls between institutional chairs within the UN system, but also within governments. The various UN entities engaged with facilitating intergovernmental talks on AI (such as the ITU, the Office for Digital and Emerging Technologies (ODET), UNESCO and the UN Secretariat) are limited in their scope of action. The approach remains siloed around certain aspects, without holistic coordination of government talks or industry efforts in a way that allows all aspects and dimensions of the problem to be addressed. Governments also struggle to overcome internal silos. Responsibility for steering national AI policy is often fragmented across different ministries and agencies. For instance, questions around how AI can and should be developed and deployed responsibly are typically championed by ministries of economy or industry, while concerns around misuse of AI fall under several thematic ministries or governmental agencies: cybersecurity authorities for cyber-related issues, ministries of the interior or justice for political issues, and ministries for foreign affairs or defence for aspects related to CBRN weapons risks. This fragmentation can lead to a lack of information sharing and coordination among different government departments, as well as competition or duplication of efforts. This is a significant challenge for such a collective action problem, though this is not the first time governments have had to coordinate internally and externally to address shared problems.

Another institutional-level issue is that states are unequally equipped to guide, support or oversee responsible innovation in their national AI community. A lot of states do not have the necessary infrastructure or human resources within government to conduct risk assessments or support responsible innovation efforts in the AI industry and academia. The launch in November 2024 of a network of national safety institutes could help states overcome such limitations through information and cooperation, but the network remains in its early stages.[118] Moreover, many, if not all, of the national safety institutes are still in an implementation phase with their mandate and missions being shaped along the way.

Finally, the evolution of the geopolitical landscape over the past five years has made it much more difficult for states and international organizations to engage with one another on AI risk management. It has also led several of them to reconsider their priorities. For example, after the Russian invasion of Ukraine, several European countries reconsidered their views on dual-use risks. The fact that the Ukrainian armed forces have been able to rely extensively on modified commercial technologies to defend themselves has made the weaponization of civilian technologies less of a concern. Germany, for instance, started reconsidering the 'civil clause' that constitutionally prohibited some German universities from working on military research or research with

[118] European Commission, 'First meeting of the International Network of AI Safety Institutes', News, 20 Nov. 2024.

a strong dual-use dimension.[119] The European Commission proposed to allow dual-use research projects in its next Framework Programme for Research and Innovation, due to start in 2028.[120] The outcome of the presidential election in the USA in 2024 also led to a major shift in the position of the US government on the regulation of AI. Through its Executive Order 14110 on Safe, Secure and Trustworthy Development and Use of AI, the previous administration had taken steps to encourage actors in the private sector to collaborate with the government on the prevention of misuse risks and other large-scale risks that could emerge from AI development and deployment. As soon as it took office, the current administration revoked the executive order and indicated that it wanted to remove 'policies and directives that act as a barrier to American AI innovation'.[121] Discourses both within and outside the US government around the strategic importance of maintaining American dominance (over China) also indicate a decreased appetite in the USA for international coordination around AI safety and security.[122]

The results of these challenges are problematic. They fuel a fragmentation of the governance landscape and expose the AI community to different, if not contradictory, signals regarding its role and responsibility in addressing the risks that AI systems pose, including in the realm of international peace and security. Moreover, they potentially create loopholes or governance gaps that could make certain risks materialize or grow in magnitude.

A way forward for the policy community

There are several steps or options that states and international organizations could explore to address the consequences of the challenges outlined above. Supporting responsible innovation practices and strengthening the prevention and mitigation risks that AI systems pose to international peace and security do not necessarily require international agreement among states nor the formulation of international agreed rules or an international AI agency, and so can begin within existing frameworks. There are many concrete actions that governments and international organizations could take to make the AI community's efforts around responsible innovation more robust and effective, and to prevent or mitigate the impact of possible governance gaps in the current geopolitical environment.

At the national level or through regional organizations like the EU, the Association of Southeast Asian Nations and the Economic Community of West African States, governments could look for means to make universities include responsible innovation practices as a critical component of AI-related curriculums. This would be an effective way to support a culture of responsibility in the AI community. Educational efforts on responsible innovation should include consideration of dual-use risks, but also the societal impacts of the development and deployment of AI, including in the realm of international peace and security. Governments could also create or further develop public infrastructure and resources for independent testing and evaluations of AI models at the national level, similar to or within the national AI safety and security institutes. Such infrastructure or resources could support AI researchers and SMEs that lack the resources or expertise to assess the risks associated with their research or products, and also serve as an independent source of information and advice for government policy makers. States' access to independent expertise is essential to ensure that

[119] Kuhrt, N., 'German industry welcomes paper on military research', *Science Business*, 11 Apr. 2024.
[120] Matthew, D., 'Universities not in favour of dual-use research', *Science Business*, 19 Sep. 2024.
[121] White House, 'Removing barriers to American leadership in artificial intelligence', Presidential Action, 23 Jan. 2025.
[122] Perlo, J., 'US rejects international AI oversight at UN General Assembly', NBC News, 27 Sep. 2025.

AI policy decisions are not steered by vested corporate interests. Through the AI safety summits and similar types of gathering, states and international organizations could also support the emergence of more detailed guidance around how AI practitioners and companies should be engaging in responsible innovation (for peace and security). Such support can be expressed in various ways. Governments that have significant domestic expertise can recommend good practices or share information about the practices of their national industry champion or AI safety institute. Governments that do not have access to expertise at the domestic level, but have financial means, can instead support research by academia and civil society on the topic. Governments could also support the establishment of working groups under the auspices of an international organization, such as the ITU.

International organizations that have a mandate to work on AI or international peace and security, or both, directly or indirectly, can also implement several supporting measures. For instance, they can facilitate multi-stakeholder conversations on the risks that AI systems present in the context of their mandate. In the intergovernmental processes within the Chemical Weapons Convention (CWC) and Biological Weapons Convention (BWC), for example, states parties and the relevant secretariats have already started facilitating expert conversations and engagement with industry around how AI can be misused to develop and deploy chemical and biological weapons. The UN General Assembly could task UNODA with supporting a member state focus on the nexus between AI and information and communication technology (ICT) in the dedicated thematic groups of the Global Mechanism on ICT Security. UNODA and other relevant UN actors could also play a coordinating role between efforts undertaken in the context of the CWC, the BWC, the CCW and the Nuclear Non-Proliferation Treaty through formation of a joint working group. The new UN Independent International Scientific Panel on AI could play a critical role in raising awareness and dialogue around the possible risks stemming from cutting-edge and foreseeable development of AI (e.g. agentic AI). International organizations can also facilitate international dialogue and coordination on testing and evaluation methodologies. The ITU is likely to play a central role, but other actors in the UN system may also need to contribute, given that some risk evaluations require specific domain expertise. Expertise on CBRN weapon-related risks, for instance, is distributed across multiple international entities (including UNODA and the International Atomic Energy Agency), and coordination in that engagement could prove extremely valuable. In addition, an entity like the ITU could support the pooling and sharing of national resources and expertise for testing and evaluations to help states that do not have the financial or human resources to establish independent testing capabilities at the national level. There is a range of options in the toolbox, as the concluding chapters highlight.

5. Key findings

Chapters 2–4 of this report provided an overview for governmental and non-governmental actors on how advances in AI in the civilian domain could present risks to international peace and security, and how such risks can be addressed through responsible innovation. This chapter summarizes the key findings of this overview.

1. Different diagnostics, same cure: responsible innovation practice can address the full range of risks

Advances in civilian AI impact international peace and security in multiple ways. AI systems can be misused for influence operations, cyberattacks and developing weapons systems. They can also inadvertently reinforce trends that undermine the foundations of sustainable peace and security. Generative AI, for instance, is contributing to the erosion of trust in public discourse and political institutions by accelerating the pollution of the information ecosystem. While expert views on the likelihood and severity of these scenarios diverge, most of these risks could be prevented or mitigated through a greater use of responsible innovation practices within the AI community. While responsible innovation is not a silver bullet, the set of practices it involves can—when properly employed—help AI practitioners and companies to identify risks, including those outside of their immediate frame of reference, and to adopt technical and procedural measures that can meaningfully reduce the likelihood or scale of a given risk's impact.

2. Responsible innovation practices within the AI community are progressing, but inconsistently

Responsible innovation practices are increasing among the AI community, including in relation to issues directly connected to international peace and security. Companies developing and deploying the most advanced AI models routinely deploy technical and procedural measures to reduce the likelihood and potential impacts of AI misuse for political, criminal and violent purposes. There is an active conversation in academia about practices that can make AI safer, more secure, more trustworthy and, ultimately, less likely to cause large-scale harm. However, progress implementing such practices has been uneven across the AI industry and academia. For instance, SMEs consulted as part of this project commonly reported that dual-use concerns were rarely a top priority and that they felt far less equipped than major AI companies to integrate and engage with these concerns in their workflows. The efforts of major AI companies have also been inconsistent over time, and companies have been unevenly transparent about their risk management methods. In academia, efforts to mainstream education and capacity building in responsible AI remain limited. Supporting materials from civil society are predominantly in English and rarely link AI innovation explicitly to international peace and security, save for issues like CBRN weapons risk.

3. Responsible innovation is a collective action problem that needs internationally coordinated governmental interventions

Self-governance within the AI community will not be sufficient to ensure international peace and security risks associated with civilian AI are identified and addressed in a timely and effective way. AI risk management is a collective action problem that requires governmental interventions and international coordination to ensure that

minimum standards are applied. The need for intervention is increasingly recognized. Recent initiatives like the EU AI Act and the G7 Code of Conduct for General Purpose AI are notable for setting baseline requirements. However, there is room for improvement at multiple levels.

First, international coordination among states on AI governance remains limited. The fragmentation of the regulatory landscape that could result from that is problematic not only because it can be difficult for AI organizations to navigate, but also because it means that governance gaps could allow certain risks to materialize.

Second, the issue of how advances in civilian AI can impact peace and security still falls between institutional chairs in the UN system. None of the various UN entities facilitating intergovernmental engagement on AI (including the ITU, ODET, UNESCO and UNODA) can approach the topic holistically and coordinate talks in a way that allows all aspects and dimensions of the problem to be addressed simultaneously.

Third, norms and standards developed and promoted so far in multilateral settings remain very general. Much work remains to be done to put these principles into practice, from greater coordination to information and resource sharing on risk assessment. Given the pace at which novel developments in AI are being deployed, it is also mission-critical for governance efforts to monitor emerging technological developments, such as agentic AI. The creation of the UN International Independent Scientific Panel on AI is a positive contribution from this perspective.

6. Recommendations

Responsible innovation is an effective methodology for addressing AI risks to international peace and security. These practices must be further promoted, adopted and harmonized. To this end, this chapter sets out specific recommendations targeted at academia, industry, states and international organizations.

To academia

AI practitioners in academia can promote responsible innovation practices through exemplarity, education and peer review.

1. Lead by example

Academics, especially tenured professors, generally enjoy flexibility regarding the topics they research, how they conduct that research, and how they share their findings. They can therefore lead by example by adopting and publicly demonstrating best practices for responsible research and innovation in their work and publications, modelling good behaviour.

2. Mainstream responsible innovation practice in STEM education

Professors and faculty can compensate for the lack of formal training in responsible innovation in STEM education by creating opportunities for students to discuss in technical classes how their research could be misused and what they could do at a technical level to address these risks. The authors of this report have proposed concrete activities for professors to use with their students in a handbook published by UNODA.[123]

3. Encourage peers

Academics can also use mechanisms like the peer-review process for publications, calls for papers, and the creation of prizes to encourage peers to pay greater attention to the downstream consequences of their work. Academic leaders in positions of authority could also make the demonstration of responsible innovation practices a criterion of career promotion for researchers and professors.

4. Foster interdisciplinary approaches

Interdisciplinary approaches to research are key to responsible innovation. Professors in technical disciplines should seek opportunities for their students to engage with other disciplines, including social sciences, law and humanities.

To industry

AI companies could strengthen responsible innovation practices in the development and deployment of AI products and services.

1. Increase transparency

AI companies could be more transparent about the methods and processes they use to assess risks, including the extent to which they rely on external domain experts and governments to evaluate risks related to international peace and security (e.g. political misuse, cybersecurity risks and weapons development).

[123] Boulanin, V., Ovink, C. and Palayer, J., 'Handbook on responsible innovation in AI for international peace and security', UNODA Occasional Paper No. 45, July 2025.

2. Improve conditions for third-party evaluation

AI companies could improve the legal and technical conditions for independent third-party evaluations, for instance by allowing vetted independent actors to interact with their models without safety restrictions.

3. Support the development of risk evaluation methods

AI companies could also more actively support the development of better testing and evaluation methodologies for LLM-based AI systems, by being more transparent with their own methods but also by sponsoring independent research and actively contributing to academic discussions on the topic.

4. Leverage existing resources

AI companies that have limited resources (e.g. SMEs), limited experience engaging on responsible innovation, or little familiarity with peace and security risks, can utilize accessible online resources (like self-assessment checklists) or turn to third-party actors for help.

5. Share knowledge and practices across industry and fields

Technical industry organizations like the Institute of Electrical and Electronics Engineers, the International Organization for Standardization, and the Frontier Model Forum could more actively support the advance of the science of AI safety evaluations (e.g. how to account for second- and third-order effects) and emergence of best practices for risk management, for instance by facilitating information sharing among and between companies and academia. They could also facilitate the transfer of lessons from fields of science and technology dedicated to safety-critical systems, and provide material resources that are tailored to the needs of SMEs.

To states

States, along with international organizations (see below), should create the conditions for more universal and consistent adoption of responsible innovation practices. Together, their role is to provide a coordinating and standard-setting function to ensure self-regulation is meaningful and effective.

1. Support education in responsible AI

States could incentivize universities to make responsible innovation practices a critical component in AI-related curriculums or remove existing obstacles to responsible innovation.

2. Support independent testing and evaluations

States could create or further develop infrastructure and resources for independent testing and evaluations, following the model of national AI safety and security institutes.

3. Support industry to implement responsible innovation

States could provide resources for AI companies, especially for SMEs, on how to engage in responsible innovation.

To international organizations

International organizations must work with states to stress the necessity of responsible innovation for international peace and security.

1. Raise awareness

International organizations could build a greater common understanding among states on the risks that advances in AI pose to international peace and security. For instance, the UN Independent International Scientific Panel on AI could be the primary vehicle for raising awareness about AI risks. This report recommends that the panel keep AI's impact on international peace and security as a standard item on its agenda.

2. Coordinate efforts and resources

International organizations could improve communications and coordinate initiatives among themselves so that efforts to address the risks are not restricted by institutional silos. They could also support the pooling and sharing of resources, especially for less privileges states that lack independent testing capabilities.

3. Facilitate international dialogue and coordination on testing and evaluation methodologies

International organizations such as the ITU could play a coordinating role and mobilize when relevant expertise exists in the different parts of the UN system.

About the authors

Dr Vincent Boulanin is Senior Researcher and Director of the SIPRI Governance of Artificial Intelligence (AI) Programme at SIPRI. He leads SIPRI's research on how to govern the impact of AI on international peace and security.

Jules Palayer is a Researcher in the SIPRI Governance of AI Programme. His research focuses on emerging technologies and international security, particularly the peace and security risks posed by developments in AI.

Charles Ovink is a Political Affairs Officer at the United Nations Office for Disarmament Affairs, where he leads the office's work on Promoting Responsible Innovation in AI for Peace and Security. He focuses on the intersection of risks stemming from new and emerging technologies, governance, peace and security, and responsible practices.

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE
Signalistgatan 9
SE-169 72 Solna, Sweden
Telephone: +46 8 655 97 00
Email: sipri@sipri.org
Internet: www.sipri.org