RESEARCH BRIEF

AI Risks Taxonomy

Paving the Path for Confidence-Building Measures

IOANA PUSCAS

About this Brief

This research brief is a short version of the UNIDIR report on risks of artificial intelligence (AI) published in October 2023.[1]

This brief, and the full report it is based upon, are part of the UNIDIR project on **Confidence-Building Measures for Artificial Intelligence**, which has two phases. The **first phase** of the project consisted of a risk-mapping exercise, which elaborated a comprehensive analysis of risks of AI in the context of international peace and security. The **second phase** aims to build on this analysis, and to convene multiple stakeholders to explore options for confidence-building measures (CBMs) for AI that are realistic and feasible, and drawing on lessons from other processes, and from inputs from diverse actors.

Notes on Scope and Structure

An underlying objective of this project, and implicitly of this phase in the research, is to provide a clear framework to understand the risks of AI in the context of international peace and security. Various understandings of risks of AI technology have been known and part of the vocabulary of international and multilateral discussions for the past years, yet a comprehensive mapping and exploration of the risks landscape has largely been absent from such deliberations.

The taxonomy developed in this study covers key and critical areas of risks of AI, both related to the technology, and to its use and effects. To advance discussions about CBMs, and to manage the risks of AI, the international community must develop shared understandings of the risks of the technology, and how they are related or mutually reinforcing. There are many ways to classify risks of AI and that task is rendered more complex as AI is a highly scalable, general-purpose technology, with uses across domains and across (weapons) systems. For the purpose of this project, this risk taxonomy accounts for risks of AI to international peace and security and outlines the main areas of vulnerabilities of the technology as well as its potential for misuse, or for escalation in conflict.

Author

Ioana Puscas (@IoanaPuscas1) is Researcher on artificial intelligence with UNIDIR's Security & Technology Programme.

[1] Full report is available here: https://unidir.org/publication/ai-and-international-security-understanding-the-risks-and-paving-the-path-for-confidence-building-measures/

Overview of the Taxonomy

This taxonomy identifies two large clusters of risks:

RISKS OF AI TECHNOLOGY

These include vulnerabilities that are inherent to how artificial intelligence systems are built and deployed, or which arise in the context of human interaction with AI-enabled systems. Three main categories of risks are identified:

SAFETY RISKS

Failures of the technology which are due to inherent limitations of AI systems as technical systems.

CYBERSECURITY RISKS

Malicious, intentional attacks on AI systems, and which can derail its functioning or outputs in some way.

HUMAN–MACHINE INTERACTION RISKS

Undesired effects or ineffective use of AI systems due to factors typically encountered in the interaction between humans and AI systems.

RISKS OF AI TO GLOBAL SECURITY

These encompass risks that AI technologies pose to global security, including risks that emanate from the uses of AI in military contexts and in weapon systems, or more broadly, in the context of AI's convergence with other technologies and domains of warfare. This cluster of risks includes:

MISCALCULATION RISKS

Uses of AI that lead to incorrect or biased interpretations of evolving operational contexts, adversary intent, or more generally, of global competition dynamics. Risks of miscalculation are not new, but AI can magnify their scope and scale.

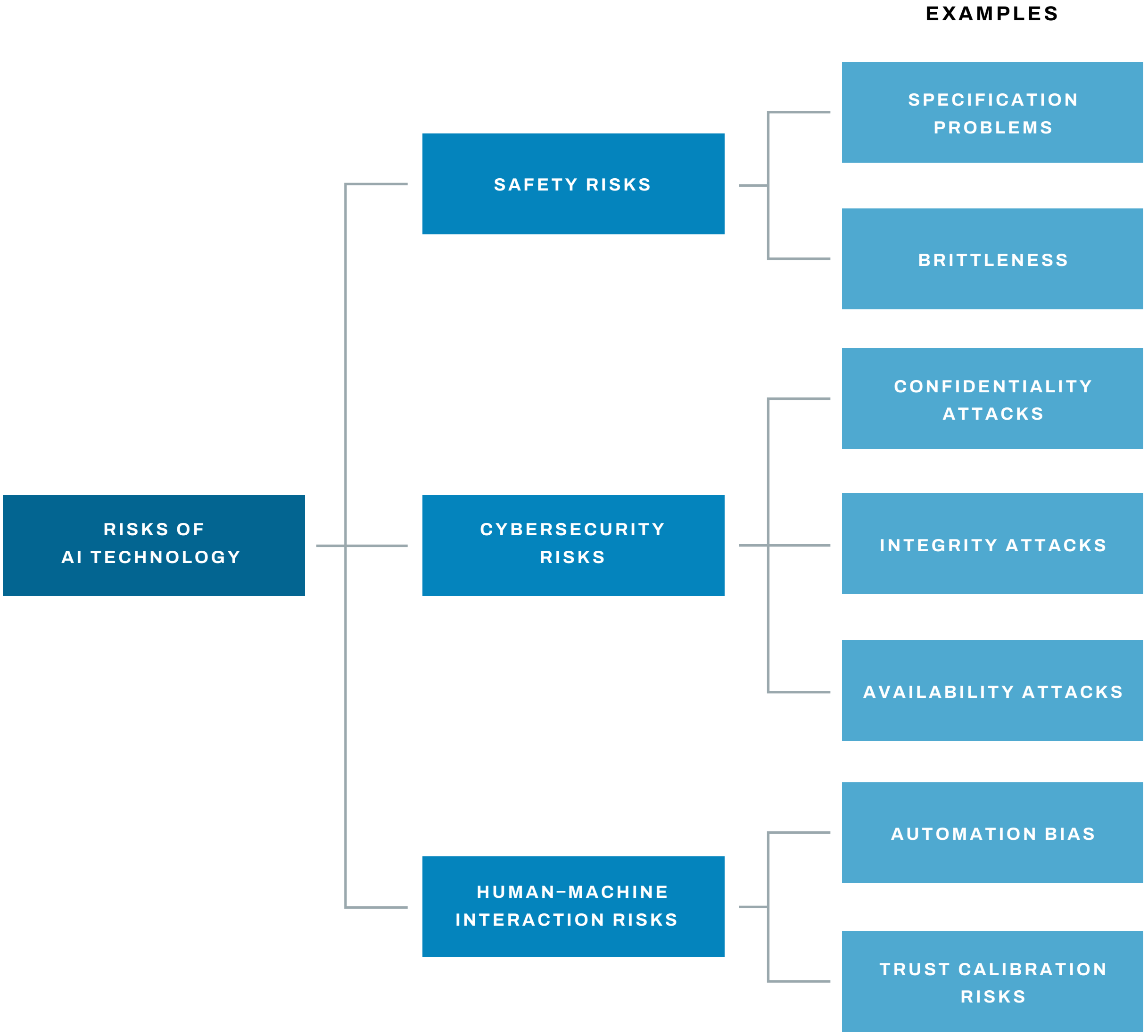
ESCALATION RISKS

AI can prompt decisions to escalate in conflict, and its potential integration into decision support or weapons systems can create direct, accidental or inadvertent forms of escalation.

PROLIFERATION RISKS

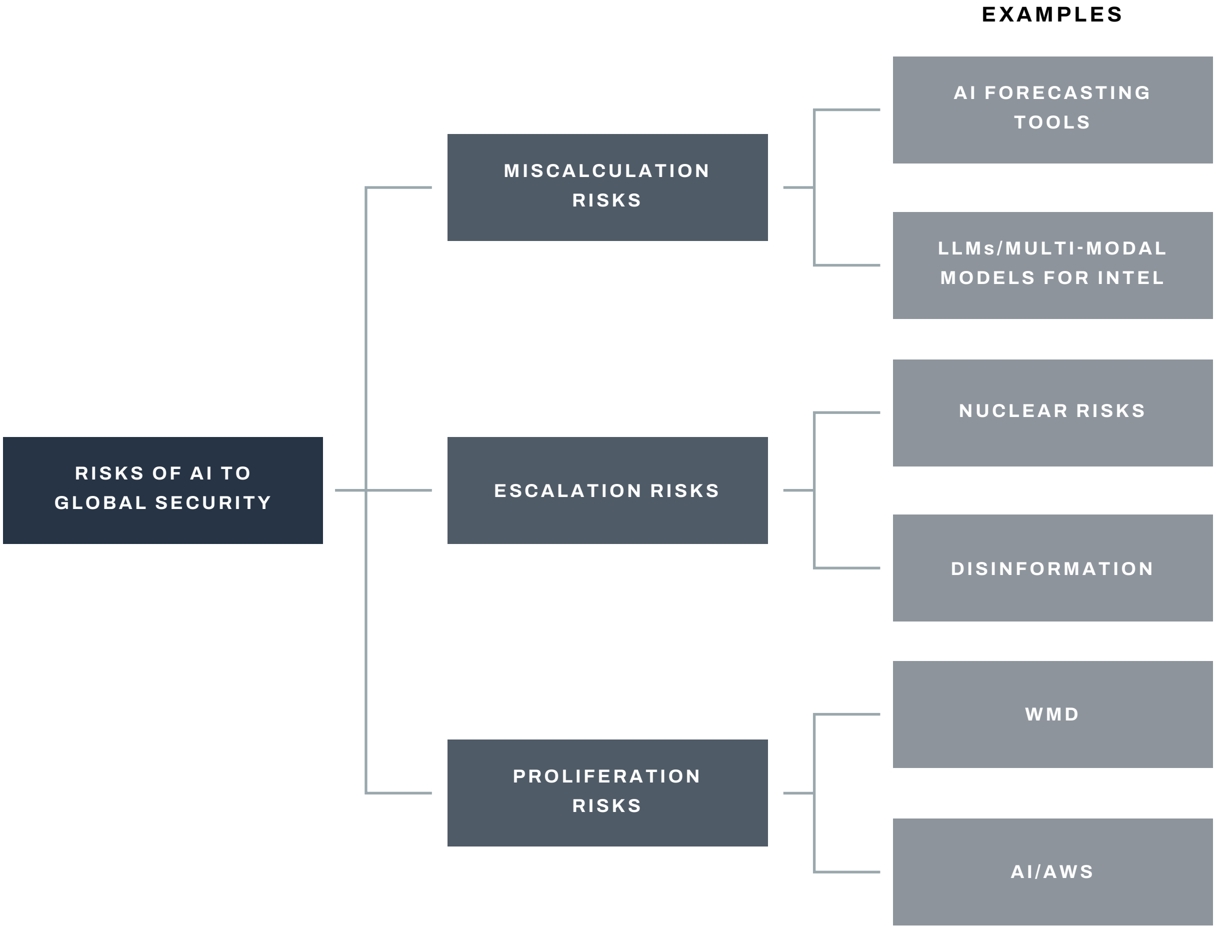
AI can alter global security dynamics and significantly enhance risks of proliferation of weapons, including weapons of mass destruction.

AI Risks Taxonomy



EXAMPLES

- RISKS OF AI TECHNOLOGY
  - SAFETY RISKS
    - SPECIFICATION PROBLEMS
    - BRITTLENESS
  - CYBERSECURITY RISKS
    - CONFIDENTIALITY ATTACKS
    - INTEGRITY ATTACKS
    - AVAILABILITY ATTACKS
  - HUMAN-MACHINE INTERACTION RISKS
    - AUTOMATION BIAS
    - TRUST CALIBRATION RISKS


EXAMPLES
RISKS OF AI TO GLOBAL SECURITY
MISCALCULATION RISKS
ESCALATION RISKS
PROLIFERATION RISKS
AI FORECASTING TOOLS
LLMs/MULTI-MODAL MODELS FOR INTEL
NUCLEAR RISKS
DISINFORMATION
WMD
AI/AWS

Part I. Risks of AI Technology

This cluster of risks includes vulnerabilities that are inherent to how artificial intelligence systems are built and deployed, or which arise in the context of human interaction with AI-enabled systems.

1. AI Safety: Inherent Risks of the Technology

Safety risks are typically defined as unintended failures of AI systems, causing them to perform incorrectly. These are inherent challenges in AI systems. Safety risks generally mean a system is not ‘attacked’ by a malicious actor, but it fails or under-performs due to one or a combination of factors.

A common issue of safety across AI applications is **brittleness**, which means that a system cannot generalize or adapt adequately to new conditions or when it is presented with new data. It is this issue that has made many AI systems at times appear highly capable in testing, only to fail dramatically when deployed under real-world conditions. Some AI systems, such as large language models (LLMs) are better at generalizing, yet that capability introduces additional safety testing requirements.

Failures in AI systems may also occur due to so-called **task specification** issues. Specification in AI refers to the task of conveying to AI systems what they need to do; in other words, to translate the intent of the designer of the system into specific actions or behaviours of the system. In practice, aligning human representation of the task to that of a machine learning (ML)[2] system is challenging, and there are many ways in which mismatches of specification can occur.

For example, the system can find ways to ‘cheat’ the specification by finding an easier method to formally complete the task but in a way that is outside of the intended goal (this is known as a ‘**reward hacking**’).

In more complex systems, such as those relying on neural networks, some specification challenges can be mitigated but other issues can emerge, such as **spurious correlations**, which are arbitrary connections between variables that are not causally related. For example, the navigation system in an uncrewed ground vehicle deployed for silent watch operations could encounter trees close to a warehouse and learn that association although the two variables are likely unrelated.

[2] ML is at the core of modern AI and the two terms are used rather interchangeably in this research.

2. Cybersecurity Risks

ML systems are vulnerable to cyberattacks typically encountered in other digital domains. These attacks are generally known by the acronym CIA—confidentiality, integrity, and availability attacks.

The table below provides an overview of the main types of attacks on AI systems.

CIA ATTACKS	WHAT IS THE GOAL?	HOW IS IT CARRIED OUT? *EXAMPLES*
Confidentiality	**Confidentiality attacks** extract hidden information about the model.	**Model extraction** – this type of attack is carried out by recording inputs and outputs of the 'victim' model until the attacker obtains enough information so that they can recreate a close copy of the stolen model. **Membership inference** – studying the inputs and outputs of the ML system to determine whether a certain data sample was part of the training dataset. **Model inversion** – attackers recover (or reconstruct) output categories of the model and try to understand key features of the input; this type of attack has been most common for image recognition systems.
Integrity	**Integrity attacks** compromise an AI system typically by altering data in some way. They have received most policy attention to date.	**Data poisoning attacks** – attackers manipulate the training dataset, which causes the system to learn counter-productively. **Evasion attacks** – malicious, extremely subtle changes to the input of the system that lead to erroneous outputs. This can be done with adversarial examples which are perturbations that cause the system to change the output (e.g., misclassify an object). In LLMs, adversarial attacks have exposed many vulnerabilities, which can steer the models towards unintended behaviours.
Availability	**Availability attacks** can lead the ML component to run slowly or completely stop.	**Sponge attacks** – in compute-intense systems, sponge examples exploit the system's dependency on hardware and soak up energy, forcing the underlying hardware to underperform.

Limitations of AI systems against cyberattacks expose vulnerabilities that are inherent to ML models as well as vulnerabilities present across the technology's life cycle. Attackers need not always break into the model itself as other opportunities for attack can be found throughout the supply chain. For example, causing a drone to misclassify targets can be done without taking control of the drone, for example by breaking into the model used by the company that developed the drone.

Attacks on ML are expected to become more frequent as the technology's uses expand, as is the range of methods used for attack. Attackers can aim to target many points of vulnerabilities, taking advantage of weaknesses in existing practices (e.g., transfer learning, which leverages trained models that are repurposed), or for example, through emerging methods such as carefully crafted camouflaging techniques specifically designed for AI systems, which can be used to mislead object detectors.

Attacks on AI systems can be relatively easy to execute, and often require less expertise than building the model in the first place. Vulnerabilities in AI systems cannot always be patched like in traditional software and although defences exist, and they can increase the cost for attackers, a persistent problem remains that while solving one set of problems, other vulnerabilities may become apparent, or the trade-offs (i.e., between safety and performance) are unacceptable. As in the cyber domain, generally, it is broadly recognized that defences for AI systems offer limited advantages before adversaries move to exploit new vulnerabilities.

3. Human–Machine Interaction Risks

Even as a system performs as expected from a technical standpoint, there are many challenges of human–machine interaction, which can lead to misuses of the technology. The way humans interact with AI systems comes with important and, in some cases unique, risks that require mitigation at various stages, including in the interface design of the system and in training requirements.

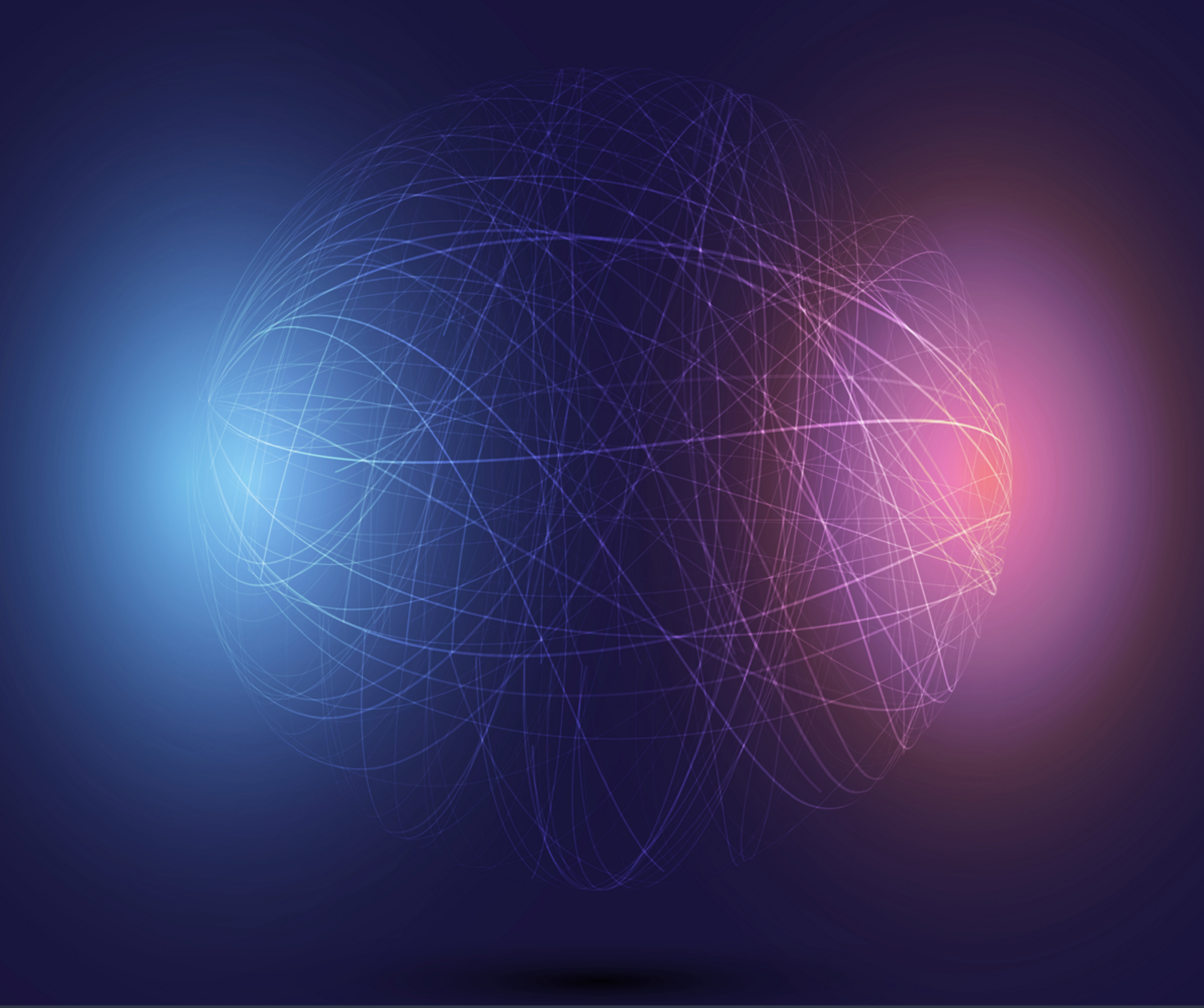
Challenges of human–machine interaction have been explored for over five decades, as automation and autonomous functions become more widely embedded in industrial processes as well as in weapons systems. The development of autonomous vehicles in recent years has further expanded the understanding of risks related to human–machine interaction in the context of complex learning systems.

A key goal in human–machine interaction is **trust calibration**, which means effectively matching a person's (e.g., operator) trust with a system's capabilities. In practice, however, trust calibration remains an enduring challenge not only in the context of human interaction with autonomous systems, but more broadly, as AI is embedded in a wide range of systems, including in intelligence analysis (explored further in Part II).

At the lower end, excessive distrust in an AI system can lead to algorithmic aversion and under-reliance on the technology. Uncritical trust, or over-trust, conversely, manifests as over-reliance on the technology, an issue also known as **automation bias**.

There are many factors that contribute to calibrating trust appropriately, ranging from individual and situational factors to **training** protocols, and **system design**, including the design of the system's interface, which plays an important role in the development of situational awareness and in the ability of the user/ operator to monitor and control a system.

Over the past decades, several incidents with autonomous systems in the military context have been attributed to problems of human–machine interaction. Operators over- or under-estimated the capabilities of the system they were operating and did not intervene at the right time or in the right manner. For example, when the US Army's Patriot missile system was involved in fratricide incidents in the early 2000s, the system did not fail, technically speaking. Several issues of human–machine interaction were at play, including operators over-trusting the system's capabilities and poorly designed interfaces.



Part II. AI and Global Security

This cluster concerns risks that AI technology introduces to global security. While the first cluster identified risks inherent to AI technology and in the context of its use, this cluster unpacks risks that AI poses or exacerbates in the context of global security.

1. Risks of Miscalculation

The incorporation of AI across intelligence, decision-support, and forecasting tools means that AI can impact human decision-making in varied ways, and with direct implications on decisions to use force. AI's potential to interfere with the entire information space raises the risk of miscalculation in warfare, though miscalculation can be more broadly conceptualized in the context of global security, including in how States assess the capabilities and behaviours of adversaries.

AI can be used at scale across **intelligence** processes, including in open-source intelligence, which makes up an overwhelming part of all intelligence activities in many States (typically, between 80 to 90 per cent). ML analytics can, for example, identify patterns in enormous data pools, or extract key information from a wide range of types of media, including images and videos. Breakthroughs in LLMs, particularly since 2022, have also been touted for use in intelligence work. LLMs may be used, for example, as assistant tools to generate summaries or to compile intelligence reports from a large body of sources. LLMs, however, are essentially probabilistic models and cannot grasp the reasoning processes that underlie intelligence work. Use of AI tools in intelligence will need careful scrutiny, both in light of the technology's current limitations (discussed in Part I) and of, quite simply, its inability at times to provide useful results.

The capabilities of LLMs are also leveraged for **battle management software tools** and to provide complex and integrated functionalities, combining intelligence collection, interactive functions and options for courses of action. While the use of AI in decision-support is not new, LLMs may fill a gap in the technology's previous shortcomings at providing a more integrated picture of the operational space.

Applications of AI in existing and emerging capabilities that directly serve to inform decision-making processes carry significant risks of miscalculation. Relative to uses of AI for autonomous weapons, the risks of using AI in intelligence and decision-support tools may appear smaller, yet the consequences can be no less devastating. Outputs of AI-powered systems can impact decisions to use force, including lethal force.

At a broader level of analysis, AI can also introduce uncertainties in international relations, globally. Developments in AI can create new perceptions of threats and vulnerabilities, and hasten the fielding of AI systems, including autonomous systems whose operation in combat may not lend itself to clear understandings of applicable rules of engagement.

2. Risks of Escalation

Escalation is a central concept in international relations, and it refers to the increase in scope and intensity of interactions between States, which can be intentional or unintentional (accidental – due to an incorrect use of a weapon, for example, or inadvertent – when an adversary acts intentionally over a threshold it considers benign, but which is unacceptable for the other side).

AI has raised significant concerns about escalation in the context of international security, including by increasing the tempo of warfare in a way that complicates de-escalation, by triggering accidents that may spiral out of control, or by prompting an escalatory chain of events even when used in less-than-lethal applications.

The use of AI in **nuclear command and control**, as well as across the **nuclear deterrence architecture**, has prompted grave concerns about AI's potential role in escalation. While it is difficult to assess the progress of AI integration in nuclear command and control, which is less likely to be swift simply because the technology is not considered predictable and robust enough, the use of AI appears promising in other areas, such as early warning systems. AI could be used to identify unusual movements or to accelerate and improve the processing of sensor data to gauge adversary behaviour. The risks remain, however, that even below firing capabilities, the use of AI might paint an incorrect picture of the adversary and raise the alert status.

AI's potential for escalation is not limited to the nuclear domain alone. AI can impact how **conventional weapons** are deployed, such as by increasing their speed and lethality, which may spiral into other pathways for escalation. For example, a State may feel legitimized to employ nuclear weapons when an adversary has gained a disproportionate advantage in its conventional forces.

Risks of escalation can also be aggravated by uses of AI to spread **disinformation**, which can cause a range of escalatory reactions in times of crisis or in combat. The use of AI to create hyper-realistic synthetic videos ('deepfakes') or other tools to disseminate deceptive or false information can complicate military campaigns, including at the tactical level, or increase the likelihood of pre-emptive attacks.

3. Risks of Proliferation

AI raises important risks of proliferation of new weapons, including weapons of mass destruction, either through AI's convergence with other domains of science and technology, or as a result of capabilities enabled by AI itself.

The **convergence** of AI with biology and chemistry holds promise for advances in the medical field, but has also exposed emerging risks of weapons proliferation. LLMs used for research and drug discovery—for example, a class of LLMs called 'chemical language models' are being used to look for new drug molecules—can also create opportunities for misuse. Further, research has demonstrated that open-source ML software can be used for the de novo design of new molecules, including toxic ones. It is, of course, important to understand that computational proof in the lab may not be very easily translated into physical weapons. However, the easing of knowledge barriers, combined with a growing number of commercial companies which provide chemical synthesis, means the risks of proliferation of **bioweapons** cannot be overlooked.

In the **cyber** domain, AI can be used to enhance the scope and scale of malicious activities in cyberspace. AI can effectively turbocharge cyber threats, and the integration of ML across technical systems means that attacks can be executed at scale. Furthermore, the risks of developing and proliferating malicious code have further magnified in the context of LLMs. Guardrails normally prevent users from requesting LLMs to generate malware, but this risk cannot be ruled out, especially as open-source models are becoming more widespread.

Finally, proliferation risks also concern the **proliferation of AI** itself and of **autonomous weapons**. AI software can be repurposed at minimal cost once developed, and it can virtually diffuse without restriction, across domains of use, and across borders.

Developments with AI technology can also enable the proliferation of autonomous weapons. The capacity of some contemporary systems to carry out a number of autonomous functions at machine speed already raises concerns about escalatory consequences.

Concerns about the proliferation of autonomous systems have grown with the surge in use of uncrewed systems over the past decades, which was also bolstered by advances in other technical domains that enabled relatively low costs of production and deployment. The breadth of risks multiplies further as more capable AI-powered systems, fitted with an expanding array of autonomous functions (including potential for operation without a human in the loop) and lethal payloads, can be fielded.

Comparative research on loitering munitions, for example, demonstrates a trend for continuous development of autonomous functions, in addition to the use of large and integrated systems. Some autonomous functions remain latent for now but that may change in light of operational needs or competitive pressures.

Conclusion

This research taxonomized the risks of AI in the context of international peace and security in two clusters: first, risks of the technology (safety, security, human–machine interaction risks), and second, risks of AI to global security (miscalculation, escalation, proliferation risks).

While it is technically and conceptually possible to present risks in different categories, it is important to highlight that risks are closely interrelated. For example, a safety failure can be exploited by malicious actors to further compromise the system, or the same vulnerability can be aggravated during use, for example if operators trust the system too much, or are inappropriately trained to understand that the system is underperforming or that it is effectively under attack.

To begin a meaningful conversation about CBMs for AI, it is critical first to map the risks of the technology and to understand their inherent complexity. Risk mitigation must be underscored by a comprehensive and informed understanding of risks. As a next step, efforts to explore options for CBMs must engage diverse stakeholders and ensure that the process is forged through their co-ownership.

UNIDIR

@unidir
/unidir
/un_disarmresearch
/unidirgeneva
/unidir

Palais des Nations
1211 Geneva, Switzerland

WWW.UNIDIR.ORG