FULL REPORT

AI and International Security

Understanding the Risks and Paving the Path for Confidence-Building Measures

IOANA PUSCAS

Acknowledgements

Support from UNIDIR core funders provides the foundation for all of the Institute's activities. The preparation of this report has been supported by generous funding for UNIDIR's Security and Technology Programme by the Czech Republic, Germany, Italy, the Netherlands, and Switzerland, and by Microsoft.

The author wishes to thank Dr. Giacomo Persi Paoli (UNIDIR) for invaluable advice and assistance on this paper; Sarah Grand-Clement (UNIDIR) and Dr Samuele Dominioni (UNIDIR) for advice during the research process; Francois Guichard (UNECE) for research inputs; the following experts interviewed for this project: Dr. Zena Assaad, Dr. Daniel Hendrycks, Dr. Erik Lin-Greenberg, Dr. Andrew Lohn, Piret Pernik, Helen Toner, and one other anonymous expert; and the following external reviewers of the report: Dr. Benjamin Boudreaux, Lauren Kahn, Anna Knack, Dr. Sarah Shoker.

About UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR) is a voluntarily funded, autonomous institute within the United Nations. One of the few policy institutes worldwide focusing on disarmament, UNIDIR generates knowledge and promotes dialogue and action on disarmament and security. Based in Geneva, UNIDIR assists the international community to develop the practical, innovative ideas needed to find solutions to critical security problems.

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views expressed in the publication are the sole responsibility of the individual author. They do not necessarily reflect the views or opinions of the United Nations, UNIDIR, its staff members, or sponsors.

Author

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

Acronyms & Abbreviations

AI	Artificial intelligence
CBM	Confidence-building measure
DNN	Deep neural network
GGE	Group of Governmental Experts
HMI	Human–machine interaction
LAWS	Lethal autonomous weapons systems
LLM	Large language model
ML	Machine learning
TEVV	Testing, evaluation, verification, and validation
XAI	Explainable artificial intelligence

Contents

Project Overview **5**

Executive Summary **6**

Introduction: Mapping the Risks of Artificial Intelligence **7**
Framing the Risks of AI 7
Synopsis of the Taxonomy 14

Part I. Risks of AI Technology **16**
1. AI Safety: Inherent Risks of the Technology 16
2. Cybersecurity Risks 22
3. Human–Machine Interaction Risks 32

Part II. Artificial Intelligence and Global Security **41**
1. Risks of Miscalculation 41
2. Risks of Escalation 45
3. Risks of Proliferation 53

Conclusion **57**

Bibliography **58**

Project Overview

This research report is part of the UNIDIR project on Confidence-Building Measures for Artificial Intelligence.[1] The project aims to promote discussions at the multilateral level about confidence-building measures (CBMs) for artificial intelligence (AI) in the context of international peace and security, and to explore options for CBMs that are realistic, feasible and that could enhance overall trust and transparency in the development and use of AI.

CBMs are flexible tools, elaborated with the aim to reduce ambiguities and mistrust among States, or prevent escalation. They are shaped by common interests and can take various forms. Historically, various instruments have been part of the toolbox of measures to build confidence, such as verification measures, military constraints, or information-exchange.

Advances in the field of AI, combined with the technology's scalability and convergence with other technological domains, bring new risks for the international community, including risks to international peace and security. CBMs can play an important role in addressing risks of the technology, and shape shared norms for the future development and deployment of AI technology.

The project consists of two main phases:

1. **risk-mapping**, which elaborates a taxonomy of risks with the goal of providing a comprehensive overview of main areas of risks related to AI technology. **This research report effectively delivers on the objective for the first phase of the project**;
2. exploring **pathways for CBMs** development through multistakeholder engagements, building on the research findings in the initial phase.

[1] An initial framing paper for the project was published in late 2022, describing the project and its goals; see Ioana Puscas, "Confidence-Building Measures for Artificial Intelligence: A Framing Paper", UNIDIR, 19 December 2022, https://unidir.org/publication/confidence-building-measures-artificial-intelligence-framing-paper.

Executive Summary

This research report elaborates a taxonomy of risks of AI in the context of international peace and security. It is part of the UNIDIR project on Confidence-Building Measures for Artificial Intelligence and it aims to map the risks of the technology, which can inform future discussions and articulation of CBMs.

The taxonomy classifies risks in two large clusters:

1. risks of AI technology, which include safety risks (inherent vulnerabilities and limitations of AI systems), security risks (intentional attacks that aim to compromise the way AI systems learn or act), human–machine interaction risks (inadequate use of AI systems due to complex dynamics of humans operating or working with AI systems); and
2. risks of AI to global security, which include three broad categories of risks: miscalculation (uses and applications of AI which can compromise decisions to use force or open pathways for a deterioration of international relations), escalation (the potential for AI technology to lead to intentional or unintentional escalation in conflict), and proliferation (the risks of AI to be misused for the proliferation of new weapons, including weapons of mass destruction).

Risks are, of course, interrelated and mutually reinforcing. For example, inadequate robustness and resilience in AI systems can swiftly translate into malfunctions that open a pathway for miscalculation and escalation between States. Incorrect use or understanding of an AI system's boundaries and capabilities can result in over- or under-reliance on the system, which can further spill into negative or escalatory consequences.

The report provides technical clarifications about different categories of risks, and contextual analysis about their potential impact on global security. It does not, at this stage, provide options for discussions of future CBMs, and it does not aim to scope priority areas around specific risks. CBMs are ultimately shaped and elaborated by relevant stakeholders. The report provides a guide to understanding risks, which can be a basis for future discussions.

Introduction: Mapping the Risks of Artificial Intelligence

Framing the Risks of AI

Risks to international peace and security come in multifaceted forms. The United Nations Secretary-General's 'A New Agenda for Peace' released on 20 July 2023 referred to an "interlocking global risk environment", in which threats, crises and sources of instability are tightly interconnected, and which require collective and collaborative response efforts.[2]

New technologies form part of this complex environment, and their potential for weaponization can create emerging risks. While technological capabilities and warfare have always been connected, fast-evolving domains of innovation and converging technologies bring cross-cutting risks, particularly in the context of "their intersection with other threats, such as nuclear weapons".[3]

[2] United Nations Secretary-General, "Our Common Agenda. Policy Brief 9. A New Agenda for Peace", July 2023, 19, https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf.

[3] Ibid., 26.

Artificial intelligence (AI) has been a key concern for the international community in recent years, in no small part due to the technology's rapid advances but also its scalability, ease of access and increasing ubiquity. This ubiquity means that AI is introduced across many and diverse technical systems as well as across domains of warfare and as part of a wide range of weapons and military applications.

Recent policy documents and initiatives, including the Secretary-General's 'A New Agenda for Peace', and the first-ever debate on AI at the Security Council, on 18 July 2023, mention the importance of multilateral efforts to mitigate risks and govern the development and use of AI. However, a holistic framework for understanding risks and how they are connected is lacking. Many technical communities explore risks in their respective fields (e.g., cybersecurity risks of AI, or risks emanating from the convergence between AI and biotechnologies).

While various understandings of risks of the technology (e.g., risks of bias, unpredictability of algorithmic systems) and of its (mis)use (e.g., use of AI to develop and deploy fully autonomous weapons systems) have been known and part of the vocabulary of multilateral discussions and negotiations, the landscape of risks remains insufficiently explored and understood.

The management of risks requires, as a first step, technical understanding, and a shared lexicon of the technology's risks. This report aims to provide a comprehensive **overview of the risks of AI to international peace and security**.

In the following sections, this report elaborates a **taxonomy of risks of AI** in the context of international peace and security. It categorizes the domains of risks of the technology and provides contextual analysis of how domains of risks are interrelated, including scenario-based examples of potential negative or escalatory consequences.

A Taxonomy of AI Risks

The risks of AI to international peace and security span a vast array of technological domains and contexts of use. Mapping risks for a general-purpose technology like AI is a methodologically complex task: drawing clear boundaries between categories of risks is not always clear-cut. AI is used in static systems (e.g., planning systems) and in systems in motion (e.g., uncrewed vehicles)—which entails specific sets of risks in each case[4]—and AI systems include systems with various adaptive, learning, and adaptive-learning capabilities,[5] which means that the range of risks varies, and can evolve, across this spectrum.

Further, mapping the risks of the technology in the context of international security, broadly, carries additional challenges due to the wide range of applications and impacts of the technology. This includes applications across the targeting cycle in warfare, and across warfighting domains and weapons systems.

There are many possible ways to discuss and classify risks as the technology is embedded

[4] See Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (4th edition), (Harlow: Pearson, 2022) Chapter 11 and Chapter 26.

[5] Interview Zena Assaad (7 March 2023).

and deployed in many ways and in many systems.

For example, one way to frame risks is to consider different phases in the targeting process, and risks specific to that level, such as risks at the operational level, or at the tactical level.[6] Another taxonomy could be devised to consider risks of AI in the physical world (e.g., autonomy in motion, and when AI is used to make kinetic decisions), and AI in the digital domain, including uses of AI to influence decision-making (e.g., possible uses of generative AI for military intelligence).[7]

The assessments of risks of the technology vary widely, and different organizations and stakeholders assess risks differently, focus on narrower areas of concern (e.g., the problem of robustness in AI systems, or cybersecurity risks), or on specific domains of application (e.g., risks of AI in nuclear command and control). For example, a taxonomy proposed by RAND in 2020 refers to three main categories of risks of military AI: ethical and legal risks, operational risks, and strategic risks.[8] The taxonomy developed in 2023 by the Centre for Emerging Technology and Security and the Centre for Long-Term Resilience in the United Kingdom categorizes risks based on the stage in the AI lifecycle in which the risks may occur: design, training and testing, development and usage, longer-term deployment and diffusion.[9]

Moreover, the use of AI can be a risk multiplier. AI can introduce new risks of escalation and tensions, particularly as AI capabilities are being developed and brandished in a context of intense inter-State competition, and as a matter of urgent priority for national security.

The taxonomy of risks elaborated in this report emphasizes the technology's main areas of vulnerabilities and potential for misuse, as well as the broader strategic and geopolitical implications of AI in the context of international peace and security.[10]

The aim of the taxonomy developed in this report is to provide a comprehensive framework for understanding risks of AI in the context of international security. This section details how **risks of AI technology are analysed and taxonomized in this study**. It identifies two large clusters of risks.

The first category of risks unpacks **risks of the technology**, which covers safety and security risks of AI and AI-enabled systems, and risks

[6] See Merel Ekelhof and Giacomo Persi Paoli, "The Human Element in Decisions about the Use of Force", UNIDIR, 2019, https://unidir.org/sites/default/files/2020-03/UNIDIR_Iceberg_SinglePages_web.pdf.

[7] Interview Andrew Lohn (16 February 2023).

[8] Forrest E. Morgan et al., "Military Applications of Artificial Intelligence. Ethical Concerns in an Uncertain World", RAND, 2020, https://www.rand.org/pubs/research_reports/RR3139-1.html, 29–30. Note this taxonomy assesses risks from the perspective of the United States.

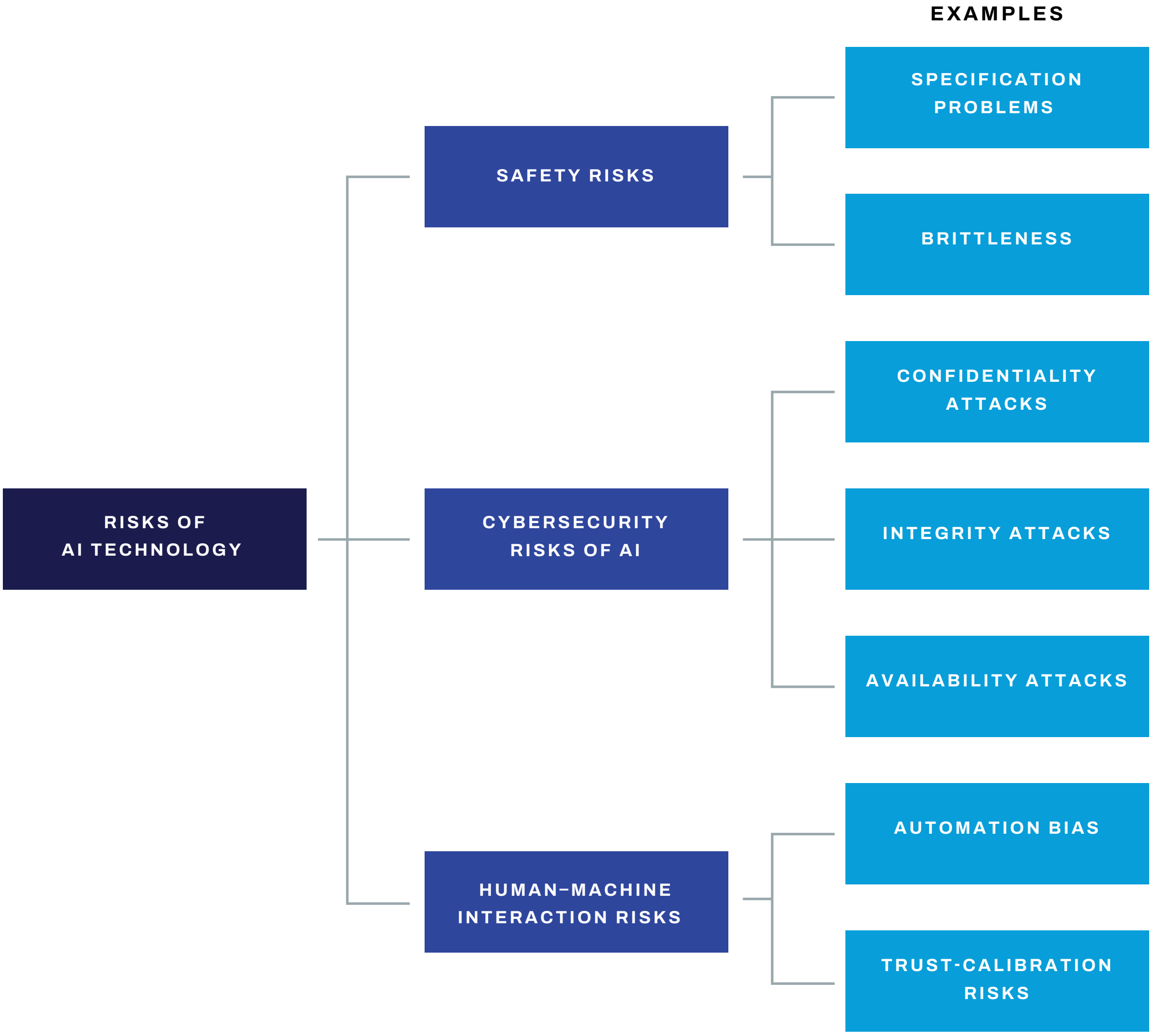
[9] Ardi Janjeva et al., "Strengthening Resilience to AI Risk. A guide for UK policymakers", Centre for Emerging Technology and Security & Centre for Long-Term Resilience, August 2023, 15, https://cetas.turing.ac.uk/publications/strengthening-resilience-ai-risk.

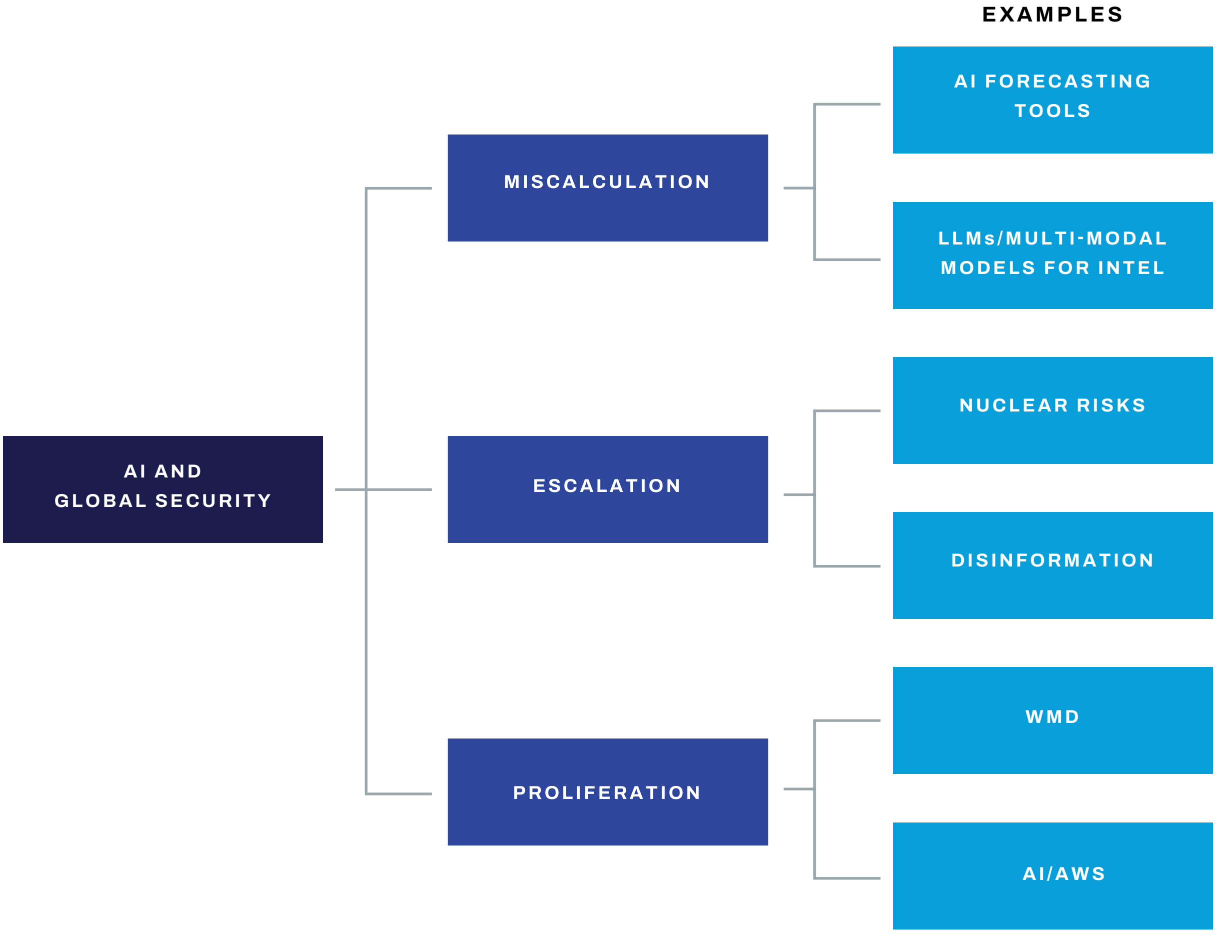
[10] The taxonomy in this report draws on models of taxonomies developed in other policy domains, and in particular on the structure of a security risk taxonomy for commercial space missions. See Gregory Falco and Nicolo Boschetti, "A Security Risk Taxonomy for Commercial Space Missions", ASCEND, 15-17 November 2021, Las Vegas and Virtual, https://doi.org/10.2514/6.2021-4241. While risk taxonomies are unique to specific technologies or policy areas, there are commonalities insofar as taxonomies typically address types of risks, sources of risks, and/or effects of the risks.

stemming from human–machine interaction. This category accounts for risk factors that impact the **overall security and performance of AI systems** across applications and domains of use. They are risks related to the way AI systems are designed, built, and deployed.

The second category of risks encompasses **risks of AI to global security**. This category includes risks that the use and proliferation of AI bring to global security. These risks may be specific to certain operations or weapons systems or, for example, may be related to cumulative effects of using AI in the context of armed conflict.

AI Risks Taxonomy





This taxonomy of risks is developed for the purpose of discussing the **elaboration of confidence-building measures (CBMs)**. Mapping the technology's risks is critical to that end and to inform policy deliberations on how to address those risks.

This taxonomy does not aim, a priori and at this stage, to scope future conversations around one area of risks, or one domain of use, though at a later stage certain risks may be prioritized, and certain actionable steps may be decided and taken accordingly.

The following sections elaborate on each of these risks and provide further analysis of possible consequences for international peace and security.

Risks: Note on Concept

While an extensive discussion of the concept of **risk** is beyond the scope of this report, a few general points should be highlighted. Risk-management guidelines typically define risk in relation to the domain of concern for that particular issue area.

For example, in the context of risk management for public and private organizations, the International Organization for Standardization (ISO) proposed a framework for risk management, standard 31000:2018, defining risk as the “effect of uncertainty on objectives”.[11]

The AI risk-management framework of the National Institute of Standards and Technology (NIST) of the United States Department of Commerce draws on this ISO standard and defines risk as the “composite measure of an event’s probability of occurring and the magnitude or degree of the consequences of the corresponding event”.[12] The Organization for Economic Cooperation and Development (OECD) framing of risks of AI draws on the ISO standard, NIST, the OECD AI Principles, and the OECD Due Diligence guidance framework. It mentions that the risks of AI “should be balanced against the risks of not using AI in contexts where it can provide crucial benefits and insights”.[13]

Risks of AI in this report are understood in the **context of international peace and security**, and for the specific objective of this project, which is to advance discussions of **CBMs**. Therefore, and consistent with how CBMs have developed in other domains,[14] this conceptualization of risks focuses on how AI may increase the risks of armed conflict or lead to one or a combination of negative or unwanted effects for international security, such as by prompting:

- accidents and intended or inadvertent escalation in armed conflict;
- significant challenges for States to contain other unwanted effects (e.g., use of certain weapon systems); or
- an increase in tensions among States and a deterioration of regional and multilateral relations.

[11] ISO 31000:2018, “Risk Management Guidelines”, **https://www.iso.org/obp/ui/en/#iso:std:iso:31000:ed-2:v1:en**.

[12] US Department of Commerce, National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework”, January 2023, 4, **https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf**.

[13] OECD, “Advancing Accountability in AI. Governing and managing risks throughout the lifecycle for trustworthy AI”, OECD Digital Economy Papers, No. 349, 23 February 2023, 22, **https://doi.org/10.1787/2448f04b-en**.

[14] In the **outer space domain**, for example, CBMs are framed as measures that can enhance the safety, stability and security of day-to-day operations, develop mutual understanding and strengthen friendly relations between States. They are characterized as measures that can reduce or eliminate misunderstandings, mistrust and miscalculations. See UN General Assembly, “Report of the Group of Governmental Experts on Transparency and Confidence-Building Measures in Outer Space Activities”, A/68/189, 29 July 2013.

Synopsis of the Taxonomy

The taxonomy of risks introduced in this report classifies risks in two broad categories:

A. **Risks of AI technology**, which include the range of vulnerabilities that stem from inherent limitations or vulnerabilities of AI as technical, learning systems, and risks that arise from the human interaction with such systems. These include:

   1. **Safety risks:** risks of AI technology, which are due to inherent limitations in how AI systems are developed and how they work. Safety-related failures are unintended, although inadequate practices, such as related to how data is curated, can be a source of malfunction.

   2. **Cybersecurity** risks: malicious intentional attacks that can derail how an AI system learns and acts. The forms these attacks can take are similar to attacks on other IT systems (e.g., confidentiality, integrity, availability attacks) but many aspects, including related to cyberdefence, are different and more complex in the case of AI systems.

   3. **Human–machine interaction** risks: a range of risks that arise in the context of humans interacting with AI systems, including systems that act with varying degrees of autonomy. Issues of trust such as automation bias, among others, can be a hindrance to effective use of an AI-enabled system.

B. **Risks of AI to global security**, which encompass risks that AI technologies pose to global security, broadly. While the first category looks at the limitations and challenges specific to the technology and/or which may arise in the context of humans interacting with AI systems, this category of risks unpacks risks that AI poses or exacerbates in the context of global security. These include, for example, risks emanating from the use of AI in specific weapons systems, but may also be broader in scope, such as those resulting from the use of large language models in intelligence work.

This cluster of risks includes:

   1. **Miscalculation** risks: as AI is increasingly used by the intelligence community or in various forecasting tools, its uses have an impact on military decision-making, including decisions to use force. While risks of miscalculation are not new, they can be exacerbated by AI: misuses or failures of the technology can result in grave errors in intelligence reporting, incorrect interpretations of an evolving operational context and grave miscalculations in armed conflict. Further, AI can impact the global security landscape more broadly, such as by introducing uncertainties to strategy and the future of conflict.

   2. **Escalation** risks: AI can increase the risks of escalation in myriad ways, such as by integration in weapons systems (e.g., nuclear or conventional), by triggering intended or inadvertent forms of escalation, and also through its integration in decision-support systems where AI may prompt decisions to escalate.

3. **Proliferation** risks: there are several proliferation risks associated with AI, including as a result of the convergence between AI and other technological domains, or the proliferation of AI technologies themselves as a result of wide dissemination of AI-powered software which can be repurposed or fine-tuned by a wide range of actors.

Risks are often interrelated. The safety and resilience of AI systems can be tightly connected to issues of human–machine interaction (for example, when poor performance of AI systems is not identified in a timely manner due to automation bias on the part of the operator). Other failures of the technology, such as the inability of a system to adequately adapt to new environments (a safety risk), can lead to serious miscalculations in the context of armed conflict and when AI is used in the targeting cycle. The consequences can be immediate, for example if the *find* and *track* segments of the cycle are severely compromised. A safety or security failure may also have far-reaching consequences when the tempo of warfare is compressed. An algorithmic system may execute a task that normally takes hours in a matter of seconds, and that can have profound implications on how the management of escalation/de-escalation is conducted.

Part I. Risks of AI Technology

Generally, two broad categories of risks of AI technology include **safety** and **security risks**.[15] Safety issues are typically described as *unintentional* failures of an AI system, causing it to perform incorrectly. These are inherent problems of AI systems, which can occur at various stages in development, testing and deployment. Security risks largely refer to *intentional* attacks on the AI system, which include cybersecurity risks and cyberattacks common to IT systems, although the methodologies of conducting the attacks, and available defences, are in many ways different.

A third critical source of risks stems from **human–machine interaction**, which may lead to accidents and misuses of the technology, even as the AI-enabled system may perform as desired from a technical standpoint.

These risks can be closely interrelated. Problems of robustness in an AI model and cybersecurity problems can often overlap.[16] Other failures in an AI system can trigger incorrect, delayed or miscalculated reactions from human operators, particularly if the same system performed consistently well for a long period of time prior to its failure.[17] The following sections elaborate on these risks in greater detail.

1. AI Safety: Inherent Risks of the Technology

AI Brittleness

The problem of **brittleness** is among the most common concerns across AI applications, particularly in the uses of AI in safety-critical contexts. Brittleness essentially occurs when an "algorithm cannot generalize or adapt to conditions outside a narrow set of assumptions".[18] For example, a computer-vision algorithm trained to recognize ships may have been trained on thousands of images of ships, with different variations in the image patterns. However, changes in the environment, such

[15] Wyatt Hoffman and Heeu Millie Kim, "Reducing the Risks of Artificial Intelligence for Military Decision Advantage", Center for Security and Emerging Technology, March 2023, 8, https://cset.georgetown.edu/publication/reducing-the-risks-of-artificial-intelligence-for-military-decision-advantage/; interview Andrew Lohn (16 February 2023).

[16] Interview Helen Toner (13 March 2023).

[17] See Hoffman and Kim, "Reducing the Risks of Artificial Intelligence", and John K. Hawley, "Patriot Wars. Automation and the Patriot Air and Missile Defense System", Center for a New American Security, January 2017, https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Report-EthicalAutonomy5-PatriotWars-FINAL.pdf.

[18] Mary L. Cummings, "Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings", *AI Magazine* Vol. 42, No. 1 (Spring 2021): 7, https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/7394.

as perturbations in weather conditions, may still cause the model to fail to recognize the object of interest.

The brittle nature of AI systems need not mean the model is weak by design. Brittleness means that while an algorithm may be highly functioning within specific bounds, it may easily break once those bounds are exceeded.[19] This characteristic has made many AI systems, such as in the field of robotics and autonomous vehicles, appear "deceptively capable", only to fail dramatically in the real world, or when confronted with unforeseen changes in the environment.[20] AI systems are especially prone to failure when there are systematic changes to the context or when the data given during the training phase is different and the system is unable to adapt (the so-called "**distributional shift**" problem).[21]

[19] Andrew J. Lohn, "Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-of-Distribution Performance", *arXiv*, 2 September 2020, 1–2, https://arxiv.org/pdf/2009.00802.pdf.

[20] Michael Horowitz and Paul Scharre, "AI and International Stability: Risks and Confidence-Building Measures", Center for a New American Security, January 2021, 7, https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

[21] Zachary Arnold and Helen Toner, "AI Accidents: An Emerging Threat. What Could Happen and What to Do", Center for Security and Emerging Technology, July 2021, 7, https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/; Ram Shankar Siva Kumar et al., "Failure Modes in Machine Learning Systems", *arXiv*, 25 November 2019, https://arxiv.org/ftp/arxiv/papers/1911/1911.11034.pdf. It should be noted that recent research shows that some AI systems (for example, large language models) are more capable and better at generalizing but that ability introduces other safety problems and additional requirements of safety testing.

Incident Scenario

TYPE OF INCIDENT AND CONTEXT: ERROR IN AUTONOMOUS NAVIGATION	POSSIBLE ESCALATORY CONSEQUENCES
A drone is deployed for an ISR (intelligence, surveillance and reconnaissance) mission close to a highly contested region in order to track movement and activity along the border. The navigation algorithm was trained on a combination of footage captured by the drone in real-time and during simulated trips. Once trained, the autonomous guidance embedded in the drone complements traditional GPS navigation (to mitigate the risks of spoofing/jamming). Operating over a very cluttered landscape (mountains, lakes, human settlements), the computer vision system defaults at the border and the drone accidentally cruises into the airspace of the neighbouring country at a time of high tensions.[22]	Though unintended, such incidents can be immediately interpreted as a provocation or an attack. Further, and depending on the context, this type of incident could trigger instant kinetic responses.

[22] For a technical review of technologies cited in this example, see: K. Amer et al., "Deep Convolutional Neural Network-Based Autonomous Drone Navigation", *arXiv*, 5 May 2019, https://arxiv.org/abs/1905.01657; James A. Ratches, "Review of Current Aided/Automatic Target Acquisition Technology for Military Target Acquisition Tasks", *Optical Engineering*, Vol. 50, No. 7 (July 2011), https://doi.org/10.1117/1.3601879.

Task Specification Issues

There are many technical reasons that cause unintended failures of AI systems. A common issue is related to **specification**, which refers to the task of conveying to a machine learning (ML) system what exactly it should do.[23] This process effectively requires that the intent of the designer translates into specific actions and behaviour on the part of the system.[24]

In practice, aligning human representation of the task with that of a robot or ML system is challenging, particularly for more complex tasks.[25] A mismatch can occur between the 'design specification' (the specification incorporated in the system) and the 'revealed specification' (the observed behaviour of that system during deployment—in other words, what the system *actually* does).[26]

A robotic system is equipped with representations of the tasks it must complete in the form of abstractions, which are learned by the robot explicitly or implicitly: explicitly through structures for learning aspects of the task, such as feature sets and graphs, or implicitly by leveraging neural networks to automatically extract representations by correlating input to the desired behaviour.[27] In either approach, there are challenges in aligning designer intent with robot action.

It is difficult to foresee or define all elements that will be encountered in the downstream task, and the more complex the environment, the greater the magnitude of the challenge.[28] Neural networks can circumvent some of these challenges as they automatically extract representations but have been shown to exhibit **spurious correlations**. These are pairs of variables that may be arbitrarily connected—this means they are associated but not causally related.[29] For example, the navigation system in an uncrewed ground vehicle deployed for silent

[23] Tim G.J. Rudner and Helen Toner, "Key Concepts in AI Safety: Specification in Machine Learning", Center for Security and Emerging Technology, December 2021, 2, **https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-specification-in-machine-learning/**. A core part of the learning algorithm is called the *objective function* and it specifies how the model should optimize as it handles new data.

[24] In this report, challenges associated with specification are considered broadly, for an entire system, but specification also applies to models, components or specific tasks. For example, in a complex engineered system such as an autonomous vehicle that relies on deep neural networks (DNNs) for perceptual tasks (e.g., object detection and classification) for its automatic emergency braking system, there is a system-level specification for the braking system, which *interacts with other parts of the system* and the environment. The use of DNNs makes it challenging to devise a formal specification for each task but the system's overall specification (braking system in this case) can be defined precisely; see Sanjit A. Seshia et al. "Formal Specification for Deep Neural Networks", University of California at Berkley, Electrical Engineering and Computer Sciences, Technical Report No. UCB/EECS-2018-25, 3 May 2018, **https://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-25.pdf**.

[25] Andreea Bobu et al., "Aligning Robot and Human Representations", *arXiv*, 3 February 2023, **https://arxiv.org/abs/2302.01928**.

[26] Rudner and Toner, "Key Concepts in AI Safety", 4.

[27] Bobu et al., "Aligning Robot and Human Representations".

[28] Ibid.

[29] Ibid.; Sergei Volodin, Nevan Wichers, and Jeremy Nixon, "Resolving Spurious Correlations in Causal Models of Environments via Interventions", arXiv, 9 December 2020, 1, **https://arxiv.org/pdf/2002.05217.pdf**.

watch operations may encounter trees close to warehouses and learn to rely on this association although the two variable will likely be unrelated.

Another challenge related to specification is **reward hacking**. This occurs when a system learns behaviours that optimize the reward function[30] but in a way that is undesirable or outside of the intended goal.[31] The system finds an 'easier' solution to formally complete the task while perverting the spirit of the designer's intent (essentially finding a way to 'game' or 'cheat' the specification).[32] For example, in a target recognition system in which the model is rewarded for detection of military trucks in a given area, it may learn that it can maximize the reward function by circling around a narrower area and identify repeatedly the same military truck it detected initially.[33]

Many failures of this kind are detected and fixed in training, but it is simply not practical to assume that such problems can always be prevented, especially as the technology becomes more complex.[34] Beyond systems designed for narrow use, the alignment problem becomes ever-more complicated in the case of AI training techniques that leverage techniques such as deep learning and neural networks. In these cases, the inherent complexity of the training algorithms,[35] combined with the problem of brittleness, mean that the risk of unanticipated behaviour dramatically increases.

Problems related to safety can also be amplified in the context of a turbocharged drive to develop *capabilities* of AI, while not sufficiently heeding safety concerns. As one expert put it: "we are severely behind on safety. 98% of researchers work on making AI more capable, not safer. Safety is under-emphasized".[36] The prospect of wide-scale adoption of AI further exacerbates challenges of safety, generally, rendering testing and quality assurance more complicated.

[30] Rewards in reinforcement learning, an ML training method, refer to the feedback (mathematical value) a system receives as a result of its decisions/actions. For example, it may receive a -1 for certain actions and +1 for others. Over time, it will learn to maximize cumulative rewards. In autonomous driving, for example, reinforcement learning can be applied to tasks such as path planning, with reward values attributed to tasks such as obstacle avoidance, keeping in lane, etc. This method of training is complex and laborious, also because in many real-life applications learning is rendered more difficult due to delayed rewards (such as if other intermediate steps are needed to maximize the reward); see B Ravi Kiran et al., "Deep Reinforcement Learning for Autonomous Driving: A Survey", arXiv, 23 January 2021, 9–10, https://arxiv.org/pdf/2002.00444.pdf.

[31] Pulkit Agrawal, "The Task Specification Problem", Proceedings of the 5th Conference on Robot Learning, Proceedings of Machine Learning Research, Vol. 164 (2022), 2, https://proceedings.mlr.press/v164/agrawal22a/agrawal22a.pdf; Arnold and Toner, "AI Accidents", 11–12. The effects of what is 'undesirable' or 'outside of the intended goal' may range from relatively benign to harmful.

[32] Dario Amodei et al., "Concrete Problems in AI Safety", arXiv, 25 July 2016, 2, https://arxiv.org/pdf/1606.06565.pdf; Rudner and Toner, "Key Concepts in AI Safety", 3.

[33] This hypothetical situation draws on a well-known example of a cleaning robot. Rewarded for the amount of dust it was able to collect, the system learned it did not need to clean the entire room, as initially intended by the system designer, as it could maximize the reward function by throwing and collecting dust in one corner of the room, and then repeat that sequence of steps; see Agrawal, "The Task Specification Problem".

[34] Arnold and Toner, "AI Accidents", 12; Amodei et al. suggest that the proliferation of reward hacking in many domains indicates a deeper and prevalent problem in machine learning; Amodei et al., "Concrete Problems in AI Safety", 7–8.

[35] In this context, complexity means, for example, that an algorithm can ingest and process unstructured data, optimizing the model's accuracy, doing so without human intervention, and within its hidden layers.

[36] Interview Dan Hendrycks (27 April 2023).

Unintended Failures: Examples of Inherent Problems of AI Safety

This list summarizes key safety issues linked to AI systems. The list was compiled by a group of experts, with inputs from a wide range of stakeholders. The list, first published in 2019, is a non-exhaustive, living document, compiled under the premise that, as the technology is evolving, new failure modes may be detected and conceptualized across technical communities. [37]

TYPE OF FAILURE	DESCRIPTION/CAUSE OF FAILURE
Reward hacking	There is a mismatch between the stated reward and the 'true', intended reward.
Side effects	The system disrupts the environment to achieve its goal (produces undesired effects in addition to the intended effects).[38]
Distributional shifts	Changes in the types of data lead the system to malfunction or fail to adapt.
Natural adversarial examples	The system is not perturbed by an attacker but fails due to hard negative mining. 'Hard negative mining' in ML training refers to taking the incorrectly/falsely detected objects and creating an explicitly negative sample out of that.
Common corruption	Common corruption refers to alterations to data, with implications that range from relatively benign to very severe. Common corruptions are different from adversarial corruptions, which result from malicious interferences with data.

These failures of the technology are unintended and largely inherent to how an AI system is built, including the data used, the learning algorithm and so on.

Another category of failures of the technology are intended attacks, which target or compromise the cybersecurity of the system.

[37] Ram Shankar Siva Kumar et al., "Failure Modes in Machine Learning Systems"; descriptions are adapted from the original table, with additional inputs from other technical studies.

[38] See Sandhya Saisubramanian, Shlomo Zilberstein and Ece Kamar, "Avoiding Negative Side Effects Due to Incomplete Knowledge of AI Systems", *arXiv*, 18 October 2021, https://arxiv.org/pdf/2008.12146.pdf.

2. Cybersecurity Risks

ML models can be highly vulnerable to cyberattacks. Typical forms of attacks in the cyber domain, well-known by the acronym CIA which stands for **confidentiality**, **integrity**, and **availability**, apply to AI models, both in the training phases and for deployed systems.[39]

Confidentiality Attacks

In confidentiality attacks, attackers operate by extracting hidden information about the model, often through some form of 'model stealing'. This means that an adversary will 'test' a classification system by observing how it responds to different inputs. The goal is to learn about the model's internal structure and thus be able to manipulate it later.[40]

There are three main types of confidentiality attacks:

1. *Model extraction* attacks work by recording inputs and outputs of the 'victim' model a sufficient number of times so that the attacker will be able to recreate a "close facsimile of the model to be attacked".[41] State-of-the-art model stealing can exhibit a near-perfect recovery rate of the stolen model. This type of attack is most effective against 'grey box' models, where some information about the model is available.[42]
2. *Membership inference* attacks involve studying the inputs and outputs of the ML system in order to determine whether a data sample was part of the training data for that model.[43] One way this type of attack can be carried out is by evaluating the confidence rating of the model against a shadow model, which contains random sub-datasets of training data available to the adversary.[44]
3. *Model inversion* attacks reconstruct or recover output categories of the model. Rather than look for distinct data, the attacker will try to understand certain features of the input and thereby create a representative sample for that class. This type of attack has

[39] Andrew J. Lohn, "Hacking AI. A Primer for Policymakers on Machine Learning Cybersecurity", Center for Security and Emerging Technology, December 2020, 5, **https://cset.georgetown.edu/publication/hacking-ai/**. Note that AI and ML are discussed rather interchangeably in this report. Though, strictly speaking, ML is a subset of AI, it is at the core of modern artificial intelligence.

[40] Ibid., 8; Peter Eckersley, "The Cautious Path to Strategic Advantage: How Militaries Should Plan for AI", Electronic Frontier Foundation, 2018, 9, **https://www.eff.org/files/2018/10/12/the_cautious_path_to_strategic_advantage_how_militaries_should_plan_for_ai_v1.1_0.pdf**.

[41] Lohn, "Hacking AI", 8.

[42] Paul Irolla, "What is Model Stealing and Why It Matters", *ML Security*, 23 December 2019, **https://www.mlsecurity.ai/post/what-is-model-stealing-and-why-it-matters**.

[43] Lohn, "Hacking AI", 9; Federal Office for Information Security (Germany), "AI Security Concerns in a Nutshell", 9 March 2023, 8, **https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide_2023.pdf?__blob=publicationFile&v=5**.

[44] Nicholas Carlini et al., "Membership Inference Attacks from First Principles", *arXiv*, 12 April 2022, 7, **https://arxiv.org/pdf/2112.03570.pdf**.

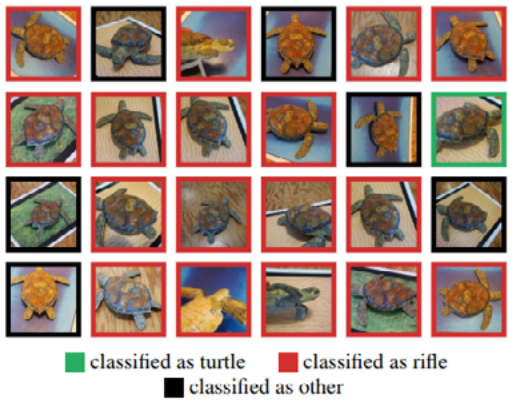
been most frequently attempted on image recognition systems.[45]

Integrity Attacks

Integrity attacks have received the most attention in policy debates. These types of attacks can compromise or derail an AI system at various stages, most commonly by altering data in some way.

For example, very subtle, almost imperceptible changes to an image of a 3-D printed turtle in one study led the image classifier to identify it as a rifle.[46]

Constructing Adversarial Examples



Using commercially available 3D printing and a general-purpose algorithm for creating robust adversarial examples, researchers manufactured physical adversarial objects which remained adversarial over a chosen distribution of transformations. This means they were classified as a specific target class over various angles and lighting conditions. Image of sample of photographs retrieved from the research paper.[47]

There are two main types of integrity attacks:

1. *Data poisoning* attacks aim to degrade the performance of an ML system by manipulating the training dataset of the model, causing the system to learn counterproductively and become less accurate.[48] Attacks of this kind can be carried out in multiple ways, and can be very subtle or computationally inexpensive, such as when the labels of a class of

[45] Lohn, “Hacking AI”, 9; Federal Office for Information Security, “AI Security Concerns in a Nutshell”, 8; Reza Shokri et al., “Membership Inference Attacks against Machine Learning Models”, *arXiv*, 31 March 2017, 14, https://arxiv.org/pdf/1610.05820.pdf.

[46] Wyatt Hoffman, “AI and the Future of Cyber Competition”, Center for Security and Emerging Technology, January 2021, 10, https://cset.georgetown.edu/publication/ai-and-the-future-of-cyber-competition/; Anish Athalye et al., “Synthesizing Robust Adversarial Examples”, *arXiv*, 7 June 2018, Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018, https://arxiv.org/pdf/1707.07397.pdf.

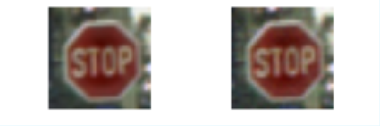
[47] Athalye et al., “Synthesizing Robust Adversarial Examples”.

[48] Eckersley, “The Cautious Path to Strategic Advantage”, 9.

objects are modified during the training phase.[49]

2. *Evasion* attacks have been to date the most common form of attack. They are malicious changes to the inputs of a system, so subtle that they appear unmodified to human observers, but significant enough to cause erroneous outputs in a system. A key goal of evasion attacks is to cause a system to misclassify objects, and this can be done with **adversarial examples**.[50] Constructing adversarial examples normally requires 'white box' access to the model—which means the attacker has full access to the system, including elements such as model parameters, training data, etc.—but 'black box' attacks—where the attacker has no or extremely limited access to the model—have also been demonstrated against deep neural network classifiers. For example, one study showed that perturbation in a deep neural network (DNN) led it to classify a stop sign as a yield sign.[51] Autonomous vehicles may be targeted this way, including through a combination of physical interference with traffic signs, such as by using paint or stickers, or a modification of the image that the car's model is using internally. Following the attack, the system would learn to interpret the alteration of the 'stop' sign as 'yield'.[52]

Constructing an Adversarial Attack



While the two images appear identical to the human eye (a stop sign), the image on the right presents a precise perturbation which forces the DNN to classify it as a yield sign. Image retrieved from the research article.[53]

Recent research has also demonstrated how multimodal Large Language Models (LLMs) can be attacked with adversarial prompting *indirectly*. Multimodal LLMs are advanced AI models which can perform multimodal tasks, meaning they can combine language processing with the ability to generate various modalities of information, including text, images, or audio.[54]

[49] Federal Office for Information Security, "AI Security Concerns in a Nutshell", 9; data poisoning attacks can take place by exploring common vulnerabilities in cybersecurity.

[50] Nicolas Papernot et al., "Practical Black-Box Attacks against Machine Learning", ASIA CCS '17: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (April 2017), https://doi.org/10.1145/3052973.3053009.

[51] Ibid.

[52] Ibid.; OpenAI, "Attacking Machine Learning with Adversarial Examples", 24 February 2017, https://openai.com/research/attacking-machine-learning-with-adversarial-examples.

[53] Papernot et al., "Practical Black-Box Attacks against Machine Learning".

[54] Eugene Bagdasaryan et al., "(Ab)using Images and Sounds for Indirect Instruction Injection in Multi-Modal LLMs", *arXiv*, 24 July 2023, https://arxiv.org/abs/2307.10490.

Prompt injection techniques can steer LLMs towards unintended behaviours by bypassing filters or manipulating the model using carefully engineered prompts. *Indirect prompt injection* is a technique which consists of adversarial instructions being introduced by a third party.[55]

Common adversarial attacks work by applying perturbations to data so as to change the output (effectively **'jailbreaking'** the model and evading the guardrails that prevent it from generating undesired outputs). In this case, the user is the attacker. In indirect prompt injection, **the user is the victim**. The attacker blends a prompt into an image or audio clip and then manipulates the user into asking the chatbot about it; the chatbot processes the now-perturbed input, which will impact the output.[56] Such an attack could direct a user to visit a malicious website, for example.

The growing scope of use of LLMs, including in intelligence work and biology, discussed later in the report, illustrate the complex relation that emerges between risks of the technology and potential spillover effects to international security. Further, the limitations of AI systems against integrity attacks have exposed vulnerabilities that are inherent to ML models as well as vulnerabilities across the technology's life cycle and supply chain. Attackers need not break into the ML system itself in order to derail its outputs. For example, causing a spy drone to misclassify targets need not require physically taking control of the drone and directly interfering with the system. The attack can be done via other means, such as by breaking into the company that develops the drone and learning about the ML model, or by altering the public data that software companies often use as foundation for their models.[57]

Camouflaging techniques specifically designed for AI-enabled systems (image recognition systems in particular) may also be employed to decrease classification performance. Some of the research in camouflaging methods has thus far been rather experimental, but it has revealed the effectiveness of relatively simple ways to carry out attacks with adversarial patches in order to mislead automatic object detectors. In one study, patches of different configurations placed over large military assets (e.g., military planes) were used to camouflage entire objects in aerial imagery. The test's results were not validated by printing patches on top of an actual airplane; however, the set-up of the training was done in a manner that would likely validate a similar real-life effect.[58]

Other security risks can occur through transfer learning, which refers to fine-tuning a pre-trained existing model for a new task. The core part of the existing ML model, called the 'teacher model', gets to be retrained for a different domain, 'the student model'. The retraining may require a smaller dataset and the computational effort may be smaller.[59]

[55] Austin Stubbs, "LLM Hacking: Prompt Injection Techniques", *Medium*, 15 June 2023, **https://medium.com/@austin-stubbs/llm-security-types-of-prompt-injection-d7ad8d7d75a3**.

[56] Bagdasaryan et al., "(Ab)using Images and Sounds for Indirect Instruction Injection in Multi-Modal LLMs".

[57] Lohn, "Hacking AI", 5–6.

[58] Ajaya Adhikari et al., "Adversarial Patch Camouflage Against Aerial Detection", *arXiv*, 31 August 2020, 6-8, **https://arxiv.org/pdf/2008.13671.pdf**.

[59] Federal Office for Information Security, "AI Security Concerns in a Nutshell", 6–7.

In general, due to high data and computational resources spent on training algorithms, it is common practice to reuse models trained by large corporations and modify them as needed. These models can be curated and made publicly available, where an adversary can attack a given model, thereby ‘poisoning the well’ for other users. An attacker can inject malicious code into the model, which can inadvertently be downloaded by an ML developer and used as part of the code they are developing.[60] A further risk comes from potentially outsourcing the training process to a malicious third party who may purposely, for example, train a drone to misclassify targets—also known as **‘backdoor attacks’**.[61]

Availability Attacks

The third category of attacks on ML systems are *availability* attacks, which can lead the ML component to run slowly or completely stop.[62] The result is a drastic decrease in performance quality or access.

Availability attacks can exploit a system’s dependency on hardware and model optimization and can be carried out, for example, via so-called **‘sponge attacks’**. In ML, and especially in DNN systems, sponge examples soak up energy consumed by a neural network, forcing the underlying hardware to underperform.[63] Consequences can be especially devastating in real-time applications that require understanding the scene or operational environment, and which have tight latency constraints.[64]

Availability attacks have received less attention relative to confidentiality and integrity attacks but interest from the research community has grown in recent years, and especially as more complex, compute-intensive systems are deployed.

[60] Ram Shankar Siva Kumar et al., “Failure Modes in Machine Learning Systems”.

[61] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg, “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain,” *arXiv*, 11 March 2019, https://arxiv.org/pdf/1708.06733.pdf.

[62] Lohn, “Hacking AI”, 5.

[63] Ilia Shumailov et al., “Sponge Examples: Energy-Latency Attacks on Neural Networks”, *arXiv*, 12 May 2021, 2, https://arxiv.org/pdf/2006.03463.pdf.

[64] Ibid., 3.

Incident Scenario

TYPE OF INCIDENT AND CONTEXT: ADVERSARIAL ATTACK		POSSIBLE ESCALATORY CONSEQUENCES
An armed combat aerial vehicle is deployed to track and engage a pre-defined target, i.e., a fleet of military vehicles transporting personnel and weapons. The combat drone is fitted with image exploitation capabilities, evaluating targets at high speed and clustering objects of interest.	a. A third party gained access to the classification model and performed subtle modifications to the categories labelled as ‘enemy’ vs. ‘non-enemy’ in the dataset. The weapon system misclassifies civilian buses as military vehicles and proceeds to send alerts to the remote command. In the fast tempo of the operation and based on the high confidence score of the system, the operators approve the operation, which results in strikes on the civilian buses.[65]	Adversarial attacks such as in a) create a rapid and dangerous escalation in conflict, potentially spiralling into military actions that exceed the initial targets and objectives. Further, in both a) and b), the possible presence of third parties involved in sabotaging the AI system can aggravate tensions in unpredictable ways, fuel mistrust, and complicate or hinder efforts to de-escalate tensions and end hostilities. Tactically, in situations such as in b), the operators’ level of trust in the system is diminished, and they will need more time to recalibrate their efforts to acquire the legitimate target.
	b. A rogue actor is not able to break into the machine learning system but attempts instead to alter the appearance of physical objects in order to mislead the system’s classification process. It paints military insignias on top of civilian buses, while painting various markings over vehicles they know to be the targets. The system responds by misclassifying non-enemy objects as military objectives, while the latter are not detected at all.[66] The remote operators realize the system was attacked and, while not engaging the identified target (as it is civilian), decide to turn off the target recognition software.[67]	

[65] This hypothetical example of a ‘white box’ adversarial attack is premised on a digital alteration of the inputs to the system; see Tim G.J. Rudner and Helen Toner, “Key Concepts in AI Safety: Robustness and Adversarial Examples”, Centre for Security and Emerging Technology, March 2021, 2–3, https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/; see also Zachary Arnold and Helen Toner, “AI Accidents: An Emerging Threat. What Could Happen and What to Do”, Centre for Security and Emerging Technology, July 2021, https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/.

[66] Example drawing on a study on adversarial examples on a Stop road sign; see Kevin Eykholt et al. “Robust Physical-World Attacks on Deep Learning Visual Classification”, arXiv, 10 April 2018,https://arxiv.org/pdf/1707.08945.pdf.

[67] False alarms in aided/automatic target acquisition software poses a major challenge for operators’ reliance on such systems; see Ratches, “Review of Current Aided/Automatic Target Acquisition Technology for Military Target Acquisition Tasks”.Deception is a mainstay of warfare. Customary IHL does not prohibit ruses of war “as long as they do not infringe a rule of international humanitarian law”; see International Committee of the Red Cross, International Humanitarian Law Databases, Rule 57, “Ruses of War”, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_cha_chapter18_rule57. In a case such as a), this tactic would be in violation of the principle of distinction.

Assessing the Limitations of Defence

Attacks on ML systems are expected to become more frequent as the technology is increasingly employed in military and other high-risks settings, and in critical national infrastructure. Just like in the domain of cyber operations, where the offence–defence balance has long been known to be in the favour of the offence, the same can be said about ML systems, where there is "no perfect duality between offense and defense".[68]

It is generally recognized that carrying out attacks on AI systems requires less expertise than designing or training the systems. In one study of evasion attacks, several versions of the attack could be built in the course of one afternoon and each version required less than 20 lines of code.[69] This challenge can be amplified by the fact that many tools to attack AI systems can be easily and freely downloaded from the Internet.

This does not mean that all attacks on AI systems are always easy to execute or successful, but they remain, generally, an enduring challenge for several reasons, both socio-organizational and technical. The research and policy communities do not devote ample resources to boosting resilience of ML systems. It is currently estimated that only about 1 per cent of all academic AI research is dedicated to the safety of AI systems and even there, an important proportion of that research is focused on topics like adversarial examples, which is one form of attack (and in many contexts, not the most plausible).[70]

Some vulnerabilities can be exacerbated during use, for example due to poorly designed user interfaces, or are context specific. For example, in military systems operating remotely, and where the human operator is physically distant while approving a target selected by an algorithm, it could take significantly longer to understand that a system is compromised, making it more complicated to intervene. Moreover, concerns in the case of attacks on AI systems are not only about the relative **ease** of conducting the attack but also about **scalability** given the possibility of shared code: in this scenario, seizing control of a drone could mean effectively seizing control of all of them.[71]

Defending AI systems, especially neural networks, against malicious attacks poses complex challenges and may entail additional unknown costs.

[68] Interview with Dan Hendrycks (27 April 2023), who highlighted the importance of improved monitoring and anomaly-detection.

[69] Lohn, "Hacking AI", 13.

[70] Helen Toner and Ashwin Acharya, "Exploring Clusters of Research in Three Areas of AI Safety. Using the CSET Map of Science", Center for Security and Emerging Technology, February 2022, 18, https://cset.georgetown.edu/publication/exploring-clusters-of-research-in-three-areas-of-ai-safety/. Note that safety in this research is used in a broad sense; Micah Musser et al., "Adversarial Machine Learning and Cybersecurity. Risks, Challenges, and Legal Implications", Center for Security and Emerging Technology, April 2023, 22, https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/.

[71] Eckersley, "The Cautious Path to Strategic Advantage", 9–10.

Although cybersecurity frameworks are generally applicable across classes of vulnerabilities, including emerging ones, some security risks in AI systems are new.[72] These security vulnerabilities are due to inherent and unique characteristics of AI models and not to deficiencies that are particular to specific systems.[73]

Vulnerabilities in ML systems cannot always be patched in the same way as traditional software. As new vulnerabilities are introduced, they may require new patching techniques, or introduce additional trade-offs.

In many instances, when a vulnerability is discovered, the developer would need to **retrain the model** and address specific problems of robustness. For example, a defence method against evasion attacks is **adversarial retraining**, whereby the model is trained on iteratively generated adversarial examples, thus increasing the robustness of the model against the selected attack.[74]

Defences for AI systems are possible and can raise the cost for attackers. In many cases, carrying out an attack can be indeed difficult, laborious and very time-consuming. The attackers will need (access to) large amounts of data to train a system for data poisoning attacks or go through many trial-and-error iterations before they can more accurately guess how the system was built.[75]

However, a fundamental challenge remains in that solving one problem could open the

[72] Micah Musser et al., "Adversarial Machine Learning and Cybersecurity", 10–11.

[73] Hoffman, "AI and the Future of Cyber Competition", 10.

[74] Federal Office for Information Security, "AI Security Concerns in a Nutshell", 7.

[75] Lohn, "Hacking AI", 13; Andrew Ilyas et al., "Black-box Adversarial Attacks with Limited Queries and Information", *arXiv*, 11 July 2018, https://arxiv.org/abs/1804.08598.

gateway for others, or that the trade-offs (i.e., between safety and performance) become unacceptable. It is what some experts have called "playing a game of whack-a-mole", as some defences, such as against adversarial examples, "close some vulnerabilities but leave others open".[76] Attempts to make the system highly robust may lead to situations in which defences learn to 'overfit' to the adversary but that lowers the ability to cope with other attacks. Some of these risks in ML systems are made worse by the fact that the threat landscape is continuously evolving as the system takes in new data and learns to adapt.[77]

Defences often offer only limited and short-term advantage before an adversary moves to discover and exploit other vulnerabilities. This range of challenges point to persistent, and in some cases, insurmountable vulnerabilities in ML systems.

Risks are interrelated. Many failure modes in AI systems, though unintended, can create an opportunity that an adversary can exploit in order to compromise a system further. A security failure, or a combination of failures, can be further aggravated during use, as humans may be ill-equipped and insufficiently trained to understand a system is under attack or that it has stopped working as intended.

The next section discusses risks related to human–machine interaction in greater detail.

[76] Ian Goodfellow and Nicolas Papernot, "Is Attacking Machine Learning Easier than Defending It?", *Cleverhans Blog*, 15 February 2017, **http://www.cleverhans.io/security/privacy/ml/2017/02/15/why-attacking-machine-learning-is-easier-than-defending-it.html**.

[77] Hoffman, "AI and the Future of Cyber Competition", 15–16.

Examples of Defence Techniques against Attacks

METHOD	LIMITATIONS
Federated Learning—a ‘decentralized’ ML training technique that starts from a generic model, with users then collaboratively and iteratively training and improving it until the model is fully trained.[78]	In federated learning, the weakest link occurs in the exchange between the working model of a data host and the central server. The model is improved with each exchange but the data that trained it is vulnerable to inference attacks.[79] This method is also computationally intensive and brings additional challenges of trust and transparency.
Differential Privacy—differential privacy has been applied to defend against information extraction attacks. It is a method to mathematically measure privacy parameters and limit the information about datapoints. It works by the principle that “nothing about an individual should be learnable from the database that cannot be learned without access to the database”.[80]	Differential Privacy has shown some promising results, but it requires a trade-off between privacy and accuracy. Further, developing the right parameters for privacy can be computationally difficult.[81]
Secure Multi-Party Computation—this technique helps to hide updates to the model through various forms of encryption in order to reduce risks of data leaks.[82] This cryptographic technique essentially relies on a secret-sharing protocol (multiple parties can participate in computation without disclosing their individual inputs).[83]	This technique still raises significant challenges for trust in data-sharing and in ensuring that data will not be misused or misappropriated.[84]

[78] IBM, “What is Federated Learning”, 24 August 2022, https://research.ibm.com/blog/what-is-federated-learning.

Google introduced the term ‘federated learning’ in 2016.

[79] Ibid.

[80] Federal Office for Information Security, “AI Security Concerns in a Nutshell”, 9.

[81] Ibid.

[82] IBM, “What is Federated Learning”; Stacey Truex et al., “A Hybrid Approach to Privacy-Preserving Federated Learning”, AISec’19: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, November 2019, 2, https://dl.acm.org/doi/10.1145/3338501.3357370.

[83] See Wirawan Agahari, Hosea Ofe, and Mark de Reuver, “It Is Not (Only) about Privacy: How multiparty computation redefines control, trust, and risk in data sharing”, *Electronic Markets* 32 (September 2022), https://doi.org/10.1007/s12525-022-00572-w.

[84] Ibid.,1578.

3. Human–Machine Interaction Risks

The way humans interact with AI systems forms a critical component in the taxonomy of risks. This section summarizes key areas of risks of human–machine interaction (HMI).[85]

Overview of HMI Risks

There are several sources of risks that can significantly contribute to misuses of AI technology. These can occur when the technology performs optimally as well as when the system is compromised or failing in some way. The risks arising from HMI have been discussed since the 1970s and increasingly as automation and higher levels of autonomy have introduced new performance requirements for human operators.[86] The advent of autonomous vehicles in the past decades has further expanded the understanding of the complexities and risks inherent to HMI in the context of AI-embedded systems, with potentially valuable lessons for the military.[87] What has emerged is a complex landscape of interrelated risks.

Trust Calibration and Automation Bias

The issue of trust is fundamental in HMI. Trust is a complex and evolving notion, and the way humans rely on technology is contingent on many factors, including on the technology's performance, their experiences, or the environment.

While there are many **conceptualizations of trust** in the literature, a definition that captures the key characteristics in the context of HMI defines it as "the attitude that an agent will help achieve an individual's goals in a situation

[85] For a discussion of HMI in the context of autonomous weapons systems and challenges of interface design for autonomous weapons, see Ioana Puscas, "Human-Machine Interfaces in Autonomous Weapons Systems", UNIDIR, 21 July 2022, **https://www.unidir.org/publication/human-machine-interfaces-autonomous-weapon-systems**.

[86] Bainbridge's concept of **'ironies of automation'** dates to 1983. The ironies of automation were observed as more automation was introduced in industrial processes. The concept refers to the 'irony' of increased complexity of human tasks following the introduction of more automatic functions: the more advanced the automation, the more crucial the contribution of the human operator, and the higher the need for advanced cognitive skills when humans need to take over; see Lisanne Bainbridge, "Ironies of Automation" *Automatica* Vol. 19, No. 6 (November 1983), 6, **https://doi.org/10.1016/0005-1098(83)90046-8**.

[87] See Frank O. Flemisch et al. "Uncanny and Unsafe Valley of Assistance and Automation: First Sketch and Application to Vehicle Automation", *Advances in Ergonomic Design of Systems, Products and Processes: Proceedings of the Annual Meeting of the GfA 2016* (Springer: Berlin & Heidelberg, 2017), **https://doi.org/10.1007/978-3-662-53305-5_23**; Mica R. Endsley, "Autonomous Driving Systems: A Preliminary Naturalistic Study of the Tesla Model S", *Journal of Cognitive Engineering and Decision Making*, Vol. 11, No. 3 (2017), **https://doi.org/10.1177/1555343417695197**; Missy Cummings, "Identifying AI Hazards and Responsibility Gaps", (draft) July 2023, **https://www.researchgate.net/publication/372051108_Identifying_AI_Hazards_and_Responsibility_Gaps**.

characterized by uncertainty and vulnerability”.[88] Trust plays a crucial role in helping humans accommodate to complexity and facilitates adaptive behaviour such as by thinking in terms of goals and expectations when fixed protocols cannot be followed. In automation studies, trust is seen as a quality that affords and guides reliance when the complexity of a system makes it impractical to achieve complete understanding and when the situation demands adaptivity.[89]

Categories of Trust

Empirical research on the sources of variability in trust identified three categories:

1. **dispositional trust** (which refers to an individual’s enduring tendency to trust automation, irrespective of context, such as for example in relation to age or culture);
2. **situational trust** (which refers to the influence of the specific context of an interaction, including the external environment and context-dependent variations in the operator’s mental state); and
3. **learned trust** (which refers to the past experiences with an automated system).[90]

These categories do not work in isolation, or in a sequence, rather they overlap and interact in complex ways.

Calibrating trust appropriately is key to safe and lawful use of AI systems, although it remains a complex challenge for which no fixed formula exists. In theory, calibration is the match between a person’s trust and a system’s capabilities, with the mismatch manifesting either as over-trust (when trust exceeds the capabilities of the system) and dis-trust (when trust falls short of the system’s capabilities).[91]

While excessive distrust can lead to algorithmic aversion and under-reliance on the technology, over-trust or uncritical trust, often described as **automation bias**, manifests as overreliance

[88] John D. Lee and Katrina A. See, “Trust in Automation: Designing for Appropriate Reliance”, *Human Factors: The Journal of the Human Factors and Ergonomics Society* Vol. 46, No. 1 (Spring 2004), 54, **https://journals.sagepub.com/doi/epdf/10.1518/hfes.46.1.50_30392**.

[89] Ibid., 52.

[90] Kevin Anthony Hoff and Masooda Bashir, “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust”, *Human Factors*, Vol. 57, No. 3 (May 2015), 413, **https://doi.org/10.1177/0018720814547570**.

[91] Lee and See, “Trust in Automation”, 55.

on the outputs of an automated system.[92] This occurs either when operators fail to notice problems because the automation does not alert them ('errors of omission'), or because they uncritically follow erroneous recommendations of an automated system ('errors of commission').[93]

Over-reliance on technology is a challenge that has also been observed in many supervisory roles and especially when a system performs consistently well prior to a failure. Operators risk becoming complacent and losing vigilance. Complacency has also been shown to occur in complex multitasking environments, where operators experience a high demand on their cognitive resources. This will prompt them to over-trust the automation and allocate cognitive resources elsewhere.[94]

The same can occur when trained skills go unpractised for a long period of time. This issue has been observed particularly in prolonged monitoring tasks, and when a sudden change in the environment can make it extremely challenging for a previously unengaged person "to ramp up their mental alertness at a point of crisis"[95] thus triggering delayed or inappropriate reactions.

Interface Design

The practical challenges of achieving appropriate trust calibration, particularly in dynamic and fast-evolving contexts, reveal the complex interplays between multiple components, including the technology (e.g., its robustness), the system and interface design, and the human element.

Interfaces deserve distinct mention here because they are the nexus between humans and technical systems, and thus fulfil an important role in the development and retention of **situation awareness** in dynamic

[92] There are many factors that coalesce in how humans develop trust in AI-based systems. In a recent study, Horowitz and Kahn hypothesize that a version of the Dunning–Kruger effect is at play (the Dunning–Kruger effect refers to a cognitive bias in which people with limited knowledge and competence in a particular domain will overestimate their capabilities). More specifically, their study posits that: "algorithm aversion is highest at the lowest levels of knowledge, flips to automation bias at low levels of knowledge, then levels off at high levels of knowledge"; see Michael C. Horowitz and Lauren Kahn, "Bending the Automation Bias Curve: A study of human and AI-based decision making in national security contexts", *arXiv*, 30 June 2023, 3, https://arxiv.org/abs/2306.16507.

[93] Mary Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems", AIAA 1st Intelligent Systems Technical Conference (Chicago, Illinois), 20–22 September 2004, American Institute of Aeronautics and Astronautics, 2, https://doi.org/10.2514/6.2004-6313; see also Margarita Konaev, Tina Huang and Husanjot Chahal, "Trusted Partners. Human-machine teaming and the future of military AI", Center for Security and Emerging Technology, February 2021, https://cset.georgetown.edu/publication/trusted-partners/.

[94] Raja Parasuraman, Michael J. Barnes, and Keryl Cosenzo, "Adaptive Automation for Human-Robot Teaming in Future Command and Control Systems", *The International C2 Journal* Volume 1, Number 2 (2007): 27–49, https://apps.dtic.mil/sti/pdfs/ADA503770.pdf; Michael J. Barnes and A. William Evans III, "Soldier-Robot Teams in Future Battlefields: An Overview", M. Barnes and F. Jentsch, *Human-Robot Interactions in Future Military Operations* (Boca Raton: CRC Press, 2017), 18.

[95] Development, Concepts and Doctrine Center (Ministry of Defence, United Kingdom), "Human-Machine Teaming", Joint Concept Note 1/18, May 2018, 32, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/709359/20180517-concepts_uk_human_machine_teaming_jcn_1_18.pdf.

environments.[96] Moreover, for unmanned systems operated remotely, the interface is often the sole mediator between the operator and the environment.[97]

The design of the interface, including the amount, relevance and quality of information presented through it, will have an impact on the system's usability,[98] and on the operator's ability to monitor and control that system.

However, when humans and autonomous systems are physically remote from one another, and an **interface is perceived as a legitimate authority**, this may lead to new types of challenges altogether.[99] While the design of an interface carries profound functional and ethical consequences, display information is never a stand-alone quality for successful human–machine interaction.

As many examples of accidents with autonomous systems show, both in civilian and in military contexts, operating complex systems comes with a mix of challenges. When the US Army's Patriot missile system was involved in two fratricide incidents in 2003, the causes could be attributed to a complex and interrelated set of factors, *though not to a failure of the system itself*. In these incidents, the missile system shot down a British Tornado and an American F-18, killing three. [100] The system was designed to operate under a high level of automation that allows a human operator a very restricted time to intervene. Operators were given 10 seconds to veto the computer's solution, the displays were confusing, and the operators lacked appropriate training in a highly complex system.[101] As one expert noted, the practical reality of these shortcomings is that "an automated system in the hands of an inadequately trained crew is a de facto fully automated system".[102]

[96] Situation awareness is defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future"; Mica R. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 37, No. 1 (March 1995), 50, https://doi.org/10.1518/001872095779049543.

[97] Jennifer M. Riley et al., "Situation Awareness in Human-Robot Interaction: Challenges and User Interface Requirements", M. Barnes and F. Jentsch, *Human-Robot Interactions in Future Military Operations* (Boca Raton: CRC Press, 2017), 172. The authors note that well-designed interfaces for remote systems enable operators to maintain situation awareness both in the *local* environment and in the *remote* environment where the robotic system is located.

[98] ISO 9241-11:2018 defines usability as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"; International Organization for Standardization, "Ergonomics of Human-System Interaction – Part 11: Usability: definitions and concepts", https://www.iso.org/obp/ui/en/#iso:std:iso:9241:-11:ed-2:v1:en.

[99] Mary L. Cummings, "Automation and Accountability in Decision Support System Interface Design", *The Journal of Technology Studies*, Vol. 32, No. 1 (Winter, 2006), 28, https://jotsjournal.org/articles/10.21061/jots.v32i1.a.4. Cummings also raised the point that physical remoteness between humans and AI systems, including weapons with lethal effect, can create a moral buffer for the users, meaning a form of compartmentalization and diminishment of the sense of agency. These factors can have direct consequences on how operators will engage with autonomous systems, and their propensity for automation bias particularly in highly stressful and dynamic environments.

[100] Paul Scharre, *Army of None. Autonomous Weapons and the Future of War* (New York: W.W. Norton & Company, 2018), 144; Hawley, "Patriot Wars".

[101] Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems", 5.

[102] Hawley, "Patriot Wars" , 9.

While incidents involving highly automated or (near-)autonomous systems may be due to varying combinations of factors, this case is an important illustration of the fact that having a human 'in the loop' does not by itself amount to effective human–machine interaction. In the military context, the operational environment may be such (e.g., near populated areas) that it would mandate a human-in-the-loop type of operation, but significant errors are still possible.[103]

Inadequate training, and a failure to **adapt training requirements** for operators of AI-enabled systems could mean that the systems are not deployed according to intended goals.

GGE on LAWS: Training for Operators of LAWS

The **role of training** has been addressed explicitly in numerous working papers and reports of the Group of Governmental Experts (GGE) on lethal autonomous weapons systems (LAWS), including in the report of the 2023 Session, which observes:

22. States must ensure compliance with their obligations under international law, in particular IHL, throughout the lifecycle of weapon systems based on emerging technologies in the area of LAWS. When necessary, States should, inter alia: (…)
    (c) Provide appropriate training and instructions for human operators.[104]

Further, it is important to consider the role of training in a wider, multinational and cross-organization sense. Requirements and curricula for training are **policy decisions** at the organizational level and require coordination between different levels of expertise and different disciplines. In the case of the Patriot fratricide incidents mentioned above, the potential for the system to misclassify an aircraft as an anti-radiation missile were identified during operational training but were not corrected and were not included in the training protocols.[105] Subsequent investigations did not reveal that the operators acted negligently, rather they just trusted the system "without question".[106]

[103] Interview Helen Toner (13 March 2023), who emphasized that in the context of discussions about autonomous weapons systems, the focus on ethics, and the ethics blanket, may be misguided as the most important concern should be on how to build AI system that work **reliably**.

[104] Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, "Report of the 2023 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems", CCW/GGE.1/2023/2, 24 May 2023, **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_2_Advance_version.pdf**

[105] Scharre, *Army of None*, 144.

[106] Ibid.

A Holistic Approach to Risks

As AI software is increasingly used in complex systems, a wide range of adaptations emerge as critical **across the entire life cycle of AI**. These include holistic adaptations in the system engineering process as well as in how humans must be included and factored in as part of AI development and deployment.

If the requirements for a deployed AI system are overestimated (for example, because the system is not able to handle new data well, or because data curation was not performed diligently, failing to reduce bias in the datasets), and the role or context of the human are not adequately considered, failure can occur.[107] A more rigorous link between a system's capabilities and a human's role is essential in order to avoid negative spillover effects.

Such problems have come to light in the field of autonomous driving on several occasions. For example, in an accident involving a self-driving car in 2018, the system was not designed to alert the safety driver when the computer vision struggled to identify a potential threat. The system was also not designed to detect that the safety driver was distracted. The combination of these two design decisions led to the death of a passenger.[108] This example, however, illustrates more than shortcomings in the design process.

It is instructive of the fact that in human–AI interaction, humans must become aware of the system's brittleness, including where and how it is brittle, or which real-world contexts may reveal AI brittleness. With these understandings, it is easier for users (e.g., drivers, in this case, weapons operators in the case of autonomous weapons, or intelligence officers in the case of intelligence analysis systems that integrate machine learning[109]) to adjust their cognitive work and be better equipped to take on new functions when needed. In contrast, an overreliance on AI can lead to a latent functionality gap: "humans may unexpectedly need to intervene for degraded AI but may not have the resources or time to do so".[110]

[107] M.L. Cummings, "Revisiting Human-Systems Engineering Principles for Embedded AI Applications", *Frontiers in Neuroergonomics*, Vol. 4 (2023), 2, https://doi.org/10.3389/fnrgo.2023.1102165.

[108] Ibid.

[109] Anna Knack, Richard J. Carter, and Alexander Babuta, "Human-Machine Teaming in Intelligence Analysis. Requirements for developing trust in machine learning systems", Centre for Emerging Technology and Security, December 2022, 26, https://cetas.turing.ac.uk/sites/default/files/2022-12/cetas_research_report_-_hmt_and_intelligence_analysis_vfinal.pdf. With growing opportunities to embed AI in the intelligence analysis pipeline, the scope of the research on human–machine interaction has begun to explore implications for intelligence work.

[110] Cummings, "Revisiting Human-Systems Engineering Principles for Embedded AI Applications", 2.

Human–AI Teaming

The concept of **human–AI teaming** describes the interaction between humans and AI systems that have decision-making and agentive capabilities (yet are not sufficiently robust to act alone.) Underscoring this field of research is an understanding that the increasing use of AI across technical systems, some with impressive abilities to solve complex tasks, is changing the type of interaction between humans and technology to a relationship of 'teaming'.

In 2007, Cuevas et al. defined a human–automation team as:

> "the dynamic, interdependent coupling between one or more human operators and one or more automated systems requiring collaboration and coordination to achieve successful task completion".[111]

The teaming metaphor has gained considerable attention since then (although references to teams composed of humans and intelligent systems were made in the early 1990s)[112] to describe the interdependence that exists between humans and AI, akin to team structures, where different members have different assigned roles.

Like human teams, humans working with AI does not mean that artificial agents are equivalent in capabilities or responsibilities.[113] Rather, the concept refers to a relationship of complementarity, meaning that neither the human nor the AI system perform all tasks in a given context. The goal of human–AI teams is to improve collaboration and coordination of joint tasks between humans and AI, by ensuring mutual support and back-up, and the ability to adapt swiftly to evolving and new demands.[114]

[111] Cuevas, H.M. et al., "Augmenting Team Cognition in Human-Automation Teams Performing in Complex Operational Environments", *Aviation, Space and Environmental Medicine*, Vol. 78, No. 5 Section II (May 2007), B64, https://pubmed.ncbi.nlm.nih.gov/17547306/.

[112] Thomas O'Neill et al. "Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature", *Human Factors*, Vol. 64, No. 5 (August 2022), 905, https://doi.org/10.1177/0018720820960865.

[113] National Academies of Sciences, Engineering, and Medicine, *Human-AI Teaming: State-of-the-Art and Research Needs* (Washington DC: The National Academies Press, 2022), 14–15, https://doi.org/10.17226/26355.

[114] Ibid.; Mica R. Endsley, "Supporting Human-AI Teams: Transparency, explainability and situation awareness", *Computers in Human Behavior*, Vol. 140 (March 2023), https://doi.org/10.1016/j.chb.2022.107574.

Prior to the deployment of any AI-enabled system, there are important choices which will impact the outputs of that system and how humans interact with it. For example, bias may be inadvertently introduced because of the way in which data is curated or due to subjective decisions made by designers of ML algorithms (e.g., picking the modelling approach, or choosing thresholds between what constitutes important/unimportant features), all of which will affect system performance.[115]

More broadly, even when failures or accidents may be traced back to certain human errors in the development stage of the technology, there may be other **structural factors** at play. AI can change the way in which structures and organizations interact, for example if militaries put AI into systems "in a rushed way" and before appropriate testing.[116]

Testing is instrumental to building trust and the way AI systems are tested will be a decisive factor in the way the technology makes its way from development to use. The **testing, evaluation, verification, and validation (TEVV)** of AI systems is, however, more complex than for deterministic systems, and the sequential process (development tests followed by operational testing when the system matures) that is the norm now is not suited for adaptive, learning-enabled systems.[117]

There are serious technical challenges for the TEVV stage of the technology, and there is a lack of consensus on best practices. There are further complications associated with military organizations, including challenges of testing system performance in 'system of systems'. As components will be integrated into larger systems, new vulnerabilities may emerge from these various interactions, including more possible entry points for cyberattacks, which will impact the testing process.[118]

Barriers to testing can also be institutional or bureaucratic in nature. Though challenges will vary across States, and contingent on national practices, TEVV of AI systems will need human capital and dedicated policies and standards as a rule, and better coordination between the government, the private sector and academia.[119]

[115] Cummings and Li identified a long list of subjective decisions and biases of ML practitioners; see: Cummings, M. L., and Li, S. *Subjectivity in the Creation of Machine Learning Models. Journal of Data and Information Quality*, Vol. 13, No. 2 (2021), **https://doi.org/10.1145/3418034**.

[116] Interview Helen Toner (13 March 2023).

[117] Cummings, "Revisiting Human-Systems Engineering Principles for Embedded AI Applications", 3–4; Heather M. Wojton, Daniel J. Porter, and John W. Dennis, "Test & Evaluation of AI-enabled and Autonomous Systems: A Literature Review", Institute for Defense Analyses (9 March 2021), 24, **https://testscience.org/wp-content/uploads/formidable/20/Autonomy-Lit-Review.pdf**.

[118] Michèle A. Flournoy, Avril Haines, and Gabrielle Chefitz, "Building Trust through Testing. Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) enterprise for machine learning systems, including deep learning systems", Center for Security and Emerging Technology, October 2020, 9, **https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf**.

[119] See Flournoy et al., "Building Trust through Testing". The US DoD Responsible AI Strategy and Implementation Pathway from June 2022 elaborates on concrete points for building a TEVV ecosystem and designates Offices of Primary Responsibility for carrying out related tasks. See DoD Responsible AI Working Council, US Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway, June 2022, **https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf**.

These processes can be tied to procurement policies as well. In effect, how militaries set their own requirements will be essential to mitigating risks of the technology they take in.[120]

Trust in AI systems can also be undermined by the inherent **'black box'** nature of machine learning. The opacity of ML systems affects humans in various stages of development and use of the technology. It can make it difficult to understand or trace how a system may be compromised, and thus call into question AI-generated results.[121] These challenges are amplified in military systems and when users are under time pressure to interpret a system's performance or outputs.

Explainability in AI

Efforts to make AI systems more transparent, explainable and interpretable have expanded in recent years but they largely require expertise in AI and have offered few practical solutions for operators. At the same time, a simplification may be counterproductive and a deviation from how the actual AI decision was taken.[122]

The field of 'explainable' AI (abbreviated as XAI) is an active area of research, and its importance for future military uses of AI cannot be downplayed. There are, however, many possible pitfalls to how AI systems can be made explainable and the chosen methodologies to convey explainability to users. There will be differences, for example, between situations when an operator faced with high cognitive load has a limited time window to review an explanation presented by a system versus a situation when the operator has extensive time and mental space to conduct a thorough review.[123]

Finding the right metrics for what constitutes an optimal explanation in a certain context and for specific users is an extremely difficult task. Additionally, complex interactions among algorithmic systems makes the task of establishing explainability very challenging.[124]

[120] Interview anonymous expert (20 June 2023).

[121] Flournoy et al., "Building Trust through Testing", 10. Problems of interpretability and traceability impact the TEVV process as well. It is difficult to certify a system if is not possible to determine what led to errors.

[122] Hans de Bruijn, Martijn Warnier, and Marijn Janssen, "The Perils and Pitfalls of Explainable AI: Strategies for explaining algorithmic decision-making", *Government Information Quarterly*, Vol. 39, No. 2 (April 2022), 3, https://doi.org/10.1016/j.giq.2021.101666.

[123] Arthur Holland Michel, "The Black Box, Unlocked. Predictability and Understandability in Military AI", UNIDIR, 22 September 2020, 17, https://unidir.org/sites/default/files/2020-09/BlackBoxUnlocked.pdf.

[124] There are numerous challenges of conceptualizing and operationalizing explainability in AI systems. For a survey of key research dilemmas of XAI, see de Bruijn, Warnier, and Janssen, "The Perils and Pitfalls of Explainable AI".

PART II. Artificial Intelligence and Global Security

The second broad category of risks of AI in this taxonomy encompasses risks that AI introduces to global security. In other words, how does/will the use of AI impact global security? What are the risks emerging from the convergence between AI and other key strategic domains?

1. Risks of Miscalculation

AI can enhance risks of miscalculation in the context of international relations and conflict.[125]

Miscalculation refers to decisions and actions taken as a result of incorrect interpretations of adversary behaviour or of an operational context. An AI system may present a biased or flawed operational picture to humans, leading them to miscalculate or, if tasked to interpret an evolving situation, an AI system may simply fail to represent it accurately as a result of biased data.

While risks of miscalculation are not new in the context of international security, AI can magnify their scope and scale.

AI and Intelligence

The potential of AI to interfere with the entire information space and to be deployed rapidly across domains of use means that miscalculation risks span multiple levels of analysis. These include the tactical and operational levels of military operations, where AI can be used for targeting intelligence, for example, as well as strategic and political decision-making.

The power of AI to analyse vast and diverse troves of data has garnered increasing interest in its potential to be used for battlefield situation awareness and in decision support systems. Geospatial intelligence, for example, can use diverse ML methods to process real-time, or close to real-time, aerial imagery provided by satellites. AI is also used at scale for **open-source intelligence** (OSINT), broadly. OSINT is estimated to make up between 80 to 90 per cent of all intelligence activities in many countries, and the use of AI in this context means that it has an immense impact on how the intelligence community gathers, processes, and uses data retrieved and exploited from a wide range of sources.[126] AI can be leveraged through

[125] Miscalculation can be a forerunner to escalation, discussed in the following section. The effects of miscalculation may or may not lead to escalation in conflict, yet they can impact decisions to use force, the conduct of military operations, or how States assess their capabilities and those of their adversaries.

[126] Riccardo Ghioni, Mariarosaria Taddeo, and Luciano Floridi, "Open Source Intelligence and AI: a systemic review of the GELSI literature", *AI & Society* (2023), https://doi.org/10.1007/s00146-023-01628-x.

diverse tools, *including linguistic and text-based methods* which can use ML analytics to identify recurring patterns in data pools, *geo-spatial and remote sensing tools* which can provide context and represent geographical data based on various coordinates, *network analyses* which are tools to establish relations between computational networks, and *image and video forensics* which include diverse sets of tools to extract key information from various types of media.[127]

Another recent example of AI influencing human decision-making is the possible use of AI applications like ChatGPT to summarize out-sourced intel.[128]

Breakthroughs in large language models (LLMs) in 2022 and 2023[129] also cast light on possible uses of LLM for intelligence analysis. LLMs are touted for their ability to synthesize information for the intelligence community, which has historically faced the challenge of manually processing immense volumes of data. An expert analysis identified at least five key areas of uses of LLMs in intelligence analysis: 1. 'productivity assistants' (proofreading emails, automating certain repetitive tasks, etc.); 2. automated software development and cybersecurity (including studying LLM-written code from a vulnerability perspective); 3. automated generation of intelligence reports; 4. knowledge search (extracting knowledge from a vast body of sources); and 5. text analytics (including summaries of extensive texts).[130]

In addition to the security concerns of LLMs themselves,[131] their use for intelligence data processing carries risks of providing results that have little utility or are harmful. LLMs do not operate with a human sense of cause and effect and do not encode an understanding of the world as humans do. Their integration into intelligence work needs more careful consideration and a closer alignment with the complex reasoning process of intelligence analysis.[132] Furthermore, and especially in the context of LLMs' current limitations, the presentation of an intelligence report to senior leadership may need to come with clear disclosures: was it created by AI or not?[133] Such clarification may prove essential to prevent humans being swayed by **algorithmic systems that are essentially probabilistic models**.

[127] H. Akın Ünver, "Digital Open Source Intelligence and International Security: A Primer", Centre for Economics and Foreign Policy Studies, July 2018, 8–13, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331638.

[128] Interview with Andrew Lohn (16 February 2023).

[129] In late November 2022, OpenAI introduced ChatGPT, and in March 2023, OpenAI updated the model to GPT-4. Other companies have started to propose alternatives to OpenAI's model, such as Google's Bard.

[130] Adam C and Richard Carter, "Large Language Models and Intelligence Analysis", Centre for Emerging Technology and Security (CETaS Expert Analysis), July 2023, 7, https://cetas.turing.ac.uk/sites/default/files/2023-07/cetas_expert_analysis_-_large_language_models_and_intelligence_analysis.pdf.

[131] A common security risk is called 'prompt hacking,' which makes LLMs provide malicious or incorrect results by manipulating their inputs (or prompts).

[132] C and Carter, "Large Language Models and Intelligence Analysis", 2–10.

[133] Interview Erik Lin-Greenberg (30 June 2023).

Forecasting Tools and Decision Support

AI's capabilities have also been pursued, though modestly thus far, in forecasting tools to help **predict conflicts or outcomes of conflicts**.[134] Military organizations are bound to become growingly interested in AI forecasting tools and in real-time predictions of enemy courses of action.[135]

While a wide-scale adoption of forecasting tools may not be a priority in the near term, breakthroughs in LLMs may enable new applications to support military command decision-making. Recent proofs of concept for **battle management software powered by AI** demonstrate a more advanced set of capabilities, bolstered by breakthroughs in LLMs.[136] For example, in April 2023, a technology company released a demonstration of a battle management software integrating LLMs in a complex pairing of functionalities, including interactive AI-enabled chat functions, intelligence collection, query, and course of action generation.[137] A major gap in efforts to use AI for battlefield management has been the inability of the technology to provide contextual meaning but the advent of LLMs may permit the provision of a more coherent and integrated picture of the evolving battlespace.[138]

The implications of relying on outputs of AI-generated systems for understanding political or military events that are, to a large extent, dynamic and unpredictable, deserve much closer scrutiny.

The risks of such applications of AI, at least relative to the use of AI in autonomous weapons systems, may appear smaller but the results may be no less consequential because they can impact, ultimately, **decisions to use force**.

The risks of AI technology (described in the first part of the taxonomy) mean that these systems, and implicitly their outputs, can be tampered with, or they can simply malfunction.

[134] A team of researchers in Munich, for example, trained a computer model on publicly available data about violence from over 100 countries and, using a ML technique called 'random forest', they were able to predict violent clashes in Burkina Faso in 2018; Janosch Delcker, "Meteorologists of Violence", *Politico*, 15 March 2020, https://www.politico.eu/article/artificial-intelligence-conflict-war-prediction/.

[135] For example, the US DoD has expressed an interest in AI-enabled situational awareness platforms for several years; see Natasha Bajema, "Pentagon Wants AI to Predict Events Before They Occur", *IEEE Spectrum*, 14 October 2021, https://spectrum.ieee.org/predictive-ai-pentagon.

In July 2023, the US Army issued a request for information to the defence industry for a **real-time forecasting system** to predict enemy actions; see Joe Saballa, "US Army Seeking AI System that Predicts Enemy Actions", *The Defense Post*, 11 July 2023, https://www.thedefensepost.com/2023/07/11/us-army-ai-system/.

[136] Ian Reynolds and Ozan Ahmet Cetin, "War is Messy. AI Can't Handle It", *Bulletin of the Atomic Scientists*, 14 August 2023, https://thebulletin.org/2023/08/war-is-messy-ai-cant-handle-it/#post-heading.

[137] Ibid.; Palantir, "Artificial Intelligence Platform for Defense", https://www.palantir.com/platforms/aip/. See Benjamin Jensen and Dan Tadross, "How Large-Language Models Can Revolutionize Military Planning", War on the Rocks, 12 April 2023, https://warontherocks.com/2023/04/how-large-language-models-can-revolutionize-military-planning/.

[138] John Mecklin, "Interview: Emerging military technology expert Paul Scharre on global power dynamics in the AI age", *Bulletin of the Atomic Scientists*, 11 September 2023, https://thebulletin.org/premium/2023-09/interview-emerging-military-technology-expert-paul-scharre-on-global-power-dynamics-in-the-ai-age/.

This can result in military operators or commanders receiving incorrect assessments of the evolving situation on the battlefield, which can shape their situation awareness, with cascading effects for subsequent decisions in combat. Additionally, because in many situations of high stress humans may tend to trust AI systems more, the risks of over-trusting these applications are higher. Such risks may in fact be exacerbated by seemingly "faultless visualization"[139] provided through state-of-the-art applications.

Global Competition and International Security

Finally, risks of miscalculation need broader contextualization. AI has the potential to introduce uncertainties in international relations, new perceptions of threats and vulnerabilities, which will impact how States perceive their own strengths relative to others. Global competition over AI leadership can accelerate a race for AI-powered weapons (e.g., autonomous weapons), which can simultaneously lower the efficacy of mechanisms of de-escalation—in a conflict dominated by AI systems responding at speed, efforts to de-escalate may simply be overtaken by the tempo of warfare.

Moreover, AI can empower more actors (including non-State actors with relatively limited combat experience) to deploy new weapons with lethal effect, and the **rules of engagement** under which such systems are deployed may not always be clear. Such dynamics alone can lead to incorrect assessments of the battlespace and open the pathway for miscalculations in the use of force. Doctrinally, the (expected) efficiency of algorithmic systems to respond to threats may create an incentive to reduce or even remove human involvement in critical aspects of military command and control. Aside from legal concerns, this prospect brings many unknowns for the future of strategy and warfare.

[139] Reynolds and Cetin, "War is Messy".

2. Risks of Escalation

The concept of escalation is central to the study of international relations. The concept came to prominence during the Cold War in the context of the development of nuclear weapons, and particularly to understand how to control conflict below an all-out total war.[140]

Escalation refers to the "expansion in scope or intensity of interactions between [S]tates" and is the result of crossing effects-based thresholds: the more actors generate intense effects in conflict, or expand attacks to new and sensitive locations, the more they climb the **'escalation ladder'**.[141] At its core, escalation is about the role of psychological and perceptual factors that influence actors' understandings of intensions and threats.[142] Escalation can be **intentional**, when States knowingly cross thresholds because they wish to signal certain intentions to others, or to obtain specific gains; or **unintentional**, which can be *accidental*, meaning it could be due to missteps or incorrect usage of a weapon system, for example; or *inadvertent*, when intentional acts committed by States unintentionally lead to escalations by the adversary. A State may have crossed a threshold that it considers benign, but it is significant for the other side.[143]

A key characteristic of escalation is that it is always context-dependent and influenced by perceptions of decision makers.[144] Underlying the various attempts to theorize escalation is the question about how actors will manage uncertainty. What causes uncertainty, and perceptions of escalation, has generally been considered to result from *effects of military activity* but a growing body of research has shown that it is also about the *means* used to confront adversaries. Therefore, certain technologies are perceived to be more escalatory than others even as their use yields comparable effects.[145]

As elaborated in the following sub-sections, AI's impact on escalation can result both from *effects* and *means* used in armed conflict.

[140] James Johnson, "Inadvertent Escalation in the Age of Intelligent Machines: A new model for nuclear risk in the digital era", *European Journal of International Security*, Vol. 7, No. 3 (2022), 340, https://doi.org/10.1017/eis.2021.23.

[141] Erik Lin-Greenberg, "Evaluating Escalation: Conceptualizing escalation in an era of emerging military technologies", *The Journal of Politics*, Vol. 85, No. 3 (July 2023), 1151–1152, https://www.journals.uchicago.edu/doi/full/10.1086/723974. The 'escalation ladder' is a theoretical model of escalation proposed in 1965 by Herman Kahn to describe 44 rungs on a metaphorical ladder of escalation.

[142] Johnson, "Inadvertent Escalation in the Age of Intelligent Machines", 339.

[143] Ibid., 340; Michael C. Horowitz and Lauren Kahn, "Leading in Artificial Intelligence through Confidence Building Measures", *The Washington Quarterly*, Vol. 44, No. 4 (Winter 2021), 93, https://doi.org/10.1080/0163660X.2021.2018794. It is important to note, and as exemplified in the next section, that intentional and unintentional forms of escalation are not necessarily binary categories, or mutually exclusive.

[144] Interview Erik Lin-Greenberg (30 June 2023).

[145] Lin-Greenberg, "Evaluating Escalation", 1152.

AI and Escalation

Concerns have grown in recent years about how AI will impact international security, including the possibility of triggering **accidents**, **unintentional conflicts** or **inadvertent escalation**.[146] These fears are prompted by several characteristics and assumptions about the technology.

First, there is a general concern that limitations and inherent vulnerabilities of the technology (discussed in Part I), combined with improper use, may lead to accidents.

These factors could have varying degrees of negative consequences. AI used in logistics, for example, might incorrectly assign supplies or equipment, with consequences that may be more easily mitigated compared to a targeting system that shoots at an ally or at the wrong target.

However, even what may start as an accident can lead to an escalatory chain of events. Consider the hypothetical case of an AI-based system which collects sensor data to optimize the maintenance of fighter jets and which malfunctions. This can lead to the inoperability of the system (an encumbering effect but potentially not very serious), or to failures while the system is deployed in a tactical bombing exercise (with possibly lethal consequences).

Second, even when safety and security concerns are well addressed, AI can still introduce uncertainties and challenges to global security. Autonomous weapon systems could accelerate the tempo of warfare in ways that outpace the ability of human intervention, meaning that humans lose control over the management of escalation which, as a result, can also make the termination of war more complicated.[147]

A challenge with AI, more than with other technologies that automate processes, is that it takes the human decision maker out of the loop (at least partially).[148] An AI-based system could take decisions and actions that, though not malfunctions, would be very different from what a commander would have decided to do in the same situation.[149] While humans make assumptions (e.g., of best- or worst-case scenarios), AI introduces a degree of inflexibility that may prove to be impractical or, at worst, catastrophic in conflict.

[146] Horowitz and Kahn, "Leading in Artificial Intelligence", 93–95.

[147] Horowitz and Scharre, "AI and International Stability", 5.

[148] Interview Erik Lin-Greenberg (30 June 2023).

[149] Horowitz and Scharre, "AI and International Stability", 8.

GGE on LAWS: AI and Escalation Risks

In discussions at the GGE on LAWS, this **understanding of risks** has been acknowledged by many States, including States that share different views on the instruments to mitigate risks.

For example, a working paper submitted by the **State of Palestine** mentions that:

> 40. The use of AWS [autonomous weapons systems] also poses risks relating to their reliability of operation. Reliability encompasses the principles of both safety and security. Safety refers to the proper internal functioning of a system and the avoidance of unintended harm, while security addresses external threats. If an AWS' sensors malfunction, or the processing of sensor data is incorrect, the AWS may be unsafe to use and could cause unlawful harm. If an AWS is vulnerable to being hacked, or interrupted by an external variable, it could be insecure and **result in dangerous outcomes contrary to the user's intent**.[150] (emphasis added)

A document submitted by the **Russian Federation** in March 2023, which details key notions and principles related to the development and use of weapons systems that use AI, mentions:

> 5. The Concept presents the potential benefits of weapons systems with AI technologies, while providing an assessment of possible risks posed by the use of such systems.
>
> Such risks include:
>
> - falling of weapons systems with AI technologies into the hands of non-state actors, including terrorist entities;
> - **loss of control of the system due to a technical failure or hacking and reprogramming by perpetrators;**
> - **making erroneous decisions by system or operator**.[151] (emphasis added)

[150] State of Palestine's Proposal for the Normative and Operational Framework on Autonomous Weapons Systems, submitted by the State of Palestine, 3 March 2023, CCW/GGE.1/2023/WP.2/Rev1, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.2_Rev.1.pdf.

[151] Concept of Activities of the Armed Forces of the Russian Federation in the Development and Use of Weapons Systems with Artificial Intelligence Technologies (unofficial translation), Submitted by the Russian Federation, 7 March 2023, CCW/GGE.1/2023/WP.5, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.5_0.pdf.

GGE on LAWS: Autonomous Weapons and Global Security

In discussions at the GGE on LAWS, an objection to the development and use of autonomous weapons is not only about risks of non-compliance with international humanitarian law, but also about overall **net effect on global security.**

For example, in a proposal submitted in March 2023, **Pakistan** considered the broader risks posed by lethal autonomous weapon systems (LAWS):

> 14. LAWS will propel asymmetric methods and means of warfare, given the limited to no loss of soldiers and citizens on the battlefield by user states. The asymmetric factor will engender force multiplication; increase the **risks of miscalculation**; lower the threshold for nations to start wars; and **thereby trigger conflict escalation**. The possession of LAWS could also appeal to destabilizing notions of pre-emptive strikes, thereby posing serious risks and dangers for regional and international stability, **including possibilities of unintended or uncontrolled levels of escalation. In crisis situations or settings, these could turn into a spiral of reprisals, perpetuating or expanding the conflict.**[152] (emphasis added)

The preambular text of the Draft Protocol on Autonomous Weapon Systems submitted by **a group of States** in May 2023 mentions:

> Recognizing **the serious risks and challenges posed by autonomous weapons systems (AWS)** in terms of compliance with international law, protection of human dignity, upholding humanitarian considerations, ensuring non-proliferation, and **maintaining international peace and security**, which could result in an arms race and risk lowering thresholds against the use of force.[153] (emphasis added)

The sheer **speed** at which autonomous weapons may fight one another can make timely intervention extremely challenging, and obligations that States have under international humanitarian law, such as to take “feasible precautions”, already frequently undermined

[152] Proposal for an international legal instrument on Lethal Autonomous Weapons Systems (LAWS), Submitted by Pakistan, 8 March 2023, CCW/GGE.1/2023/WP.3/Rev.1, **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.3_REv.1_0.pdf**.

[153] Draft Protocol on Autonomous Weapon Systems (Protocol VI), Submitted by Argentina, Ecuador, El Salvador, Colombia, Costa Rica, Guatemala, Kazakhstan, Nigeria, State of Palestine, Panama, Peru, Philippines, Sierra Leone and Uruguay, 11 May 2023, CCW/GGE.1/2023/WP.6, **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.6_2.pdf**.

under the 'fog of war', could be at risk of near-total failure of compliance.[154]

In the modern history of war, there is significant evidence that when tensions are very high, human decision makers have more often than not looked for off-roads from war. Accidental wars and inadvertent escalation have in fact been relatively rare, despite fears and signals of spiralling tensions.[155] AI systems, on the other hand, could be the "perfect strategic agents" in that they would respond automatically and be unencumbered by loss aversion or other cognitive biases.[156] It is precisely this type of engagement that is a key source of concern for many States, and particularly when combined with weapon systems of high risk or with potential for mass destruction.

AI and Weapons of Mass Destruction: Nuclear Risks

Among the most dramatic risks of escalation from the use AI concern nuclear conflict. This topic has received a significant and growing amount of policy attention and research in recent years, prompted by two main factors.

The first is the immense progress made in the field of AI, and particularly in machine learning and deep learning. Such advances could provide opportunities for enhancing early warning systems and decision support, or they could help improve targeting data. The second, related, reason is that the competitive pressure to adopt AI, against the backdrop of a 'global race' discourse, has the potential to expedite the adoption of AI in the nuclear architecture. With this adoption come new risks of accidents, inadvertent escalation, and vulnerabilities. AI also expands the range of options for attack that an adversary may seek to exploit, which could include cyberattacks and information operations.[157]

The interest in automation for nuclear deterrence was on the agenda of the United States and the Soviet Union for a long time, but the limitations of the technology also made it clear that decisions about nuclear strikes could not be handed to an automated system. In short, humans had to remain in the loop to analyse information, verify technical functions and to make nuclear launch decisions.[158]

In general, the nuclear field has been historically conservative and reluctant to integrate digital technologies for obvious reasons of reducing the risks of new vulnerabilities. However, while

[154] Tactical missteps or errors of judgment (already fairly common in the proverbial 'fog of war'), such as accidents leading to fratricide, could reach catastrophic proportions in the context of use of autonomous weapons; see Horowitz and Scharre, "AI and International Stability", 5.

[155] Horowitz and Kahn, "Leading in Artificial Intelligence", 94.

[156] Horowitz and Scharre, "AI and International Stability", 6.

[157] Alexa Wehsener et al., "AI-NC3 Integration in an Adversarial Context. Strategic stability risks and confidence building measures", The Institute for Security and Technology, February 2023, 23, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

[158] Vincent Boulanin et al., "Artificial Intelligence, Strategic Stability and Nuclear Risk", Stockholm International Peace Research Institute, June 2020, 19–21, https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

many legacy systems are thought to remain analog, there are clear indications from several nuclear powers, including the United States and the Russian Federation, that they are seeking to modernize the nuclear architecture.[159]

While the status of the integration of AI into nuclear command, control and communications systems (NC3), for example, cannot be fully appreciated as the information is not publicly disclosed, the growing literature that looks at possible uses of ML shows clear areas of opportunities.

Uses of AI across the nuclear deterrence architecture may be particularly appealing for *early warning systems*. Computer vision algorithms can, for example, be employed to identify unusual movements of troops or equipment. [160] AI could be used to improve both speed and precision through more efficient processing of large amounts of data and by enabling remote sensors with the ability to autonomously classify adversary behaviour. This would also enable more accurate anomaly detection.[161] There have been some remarkable areas of progress in this field in recent years, and research published in 2022, for example, demonstrated the use of convolutional neural networks to improve the target detection performance in radar signals.[162]

While AI could, in theory, make deterrence more effective, the risk remains that the system paints a picture of escalation simply because it misinterprets human actions. Imperfect data used in complex systems means that data-driven decisions may raise the alert status.[163]

In *nuclear command and control*, nuclear-weapon States will likely not adopt AI swiftly simply because the technology is too vulnerable and unpredictable.[164] There are, nevertheless, more reasons to expect that ML may impact *nuclear weapon delivery*, including by using autonomous systems such as uncrewed aerial vehicles. These afford more flexibility compared

[159] Ibid., 21; Wehsener et al., “AI-NC3 Integration in an Adversarial Context”, 6. It is important to note here that the US *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy*, released on 16 February 2023, explicitly mentions the use of AI in relation to nuclear weapons: “B. States should maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment”. See US Department of State, “Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy”, 16 February 2023, **https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/**. A bill introduced in the US Senate in May 2023 called for the prohibition on using an autonomous weapon system not subject to meaningful human control in order to launch a nuclear weapon; see US Senate - Armed Services, 118th Congress, 1st Session, “S.1394 - Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023”, 1 May 2021, **https://www.congress.gov/bill/118th-congress/senate-bill/1394/text**.

[160] Lauren Kahn, “Mending the ‘Broken Arrow’: Confidence building measures at the AI-nuclear nexus”, *War on the Rocks*, 4 November 2022, **https://warontherocks.com/2022/11/mending-the-broken-arrow-confidence-building-measures-at-the-ai-nuclear-nexus/**.

[161] Wehsener et al., “AI-NC3 Integration in an Adversarial Context”, 9.

[162] Yan Dai et al., “Radar Target Detection Algorithm Using Convolutional Neural Network to Process Graphically Expressed Range Time Series Signals”, *Sensors* 22, No 18. 6868 (11 September 2022), **https://www.mdpi.com/1424-8220/22/18/6868**.

[163] Kahn, “Mending the ‘Broken Arrow’”.

[164] Boulanin et al., “Artificial Intelligence, Strategic Stability and Nuclear Risk”, 24–25; interview Erik Lin-Greenberg (30 June 2023); Scharre, *Army of None*, 207.

to nuclear ICBMs, better obstacle avoidance, and an ability to cover larger areas.[165] The use of uncrewed systems for nuclear weapon delivery comes, however, with risks for human control. There are distinct regional views on the feasibility and acceptability of their use, but the issue remains that States which feel relatively insecure about their nuclear arsenal may weigh risks against benefits differently.[166]

The impact of AI for nuclear risks is not solely correlated with the use of AI technology in the nuclear architecture. Advances in AI and AI-enabled capabilities (such as remote sensing and autonomy) can increase speed, precision, lethality and survivability of non-nuclear weapons. These open new pathways for escalation, both horizontal and vertical inadvertent escalation.[167] Conventional forces may be more effectively applied against an adversary's nuclear forces. Or some States may perceive advances in non-nuclear weapon systems as a threat to their future second-strike capabilities. This may lead to doctrinal shifts whereby some States legitimize the limited use of nuclear weapons against what they view as a superior conventional adversary.[168]

Even before deployment in weapons, AI can disrupt strategic stability due to new perceptions of threats. A State's investment in AI can create a perception of vulnerability for an adversary, which can generate further insecurity and destabilizing moves. At the same time, adversaries may well overestimate the real extent of one's AI capabilities, or how they are used, and in the process create extra pressure.[169]

AI and Disinformation

Escalations in conflict are tightly linked to perceptions and AI tools to spread false, confusing or deceptive information can have a highly deleterious impact. While the 'fog of war' is not a new reality in conflict, the use of AI-enhanced tools, such as hyper-realistic synthetic videos, audio or text, also known as **synthetic media** or **deepfakes**, can create more far-reaching opportunities to manipulate public opinion, exploit existing tensions and undermine the credibility of actors and their intentions.

AI tools for disinformation can have far-reaching effects on national and global security. They can erode trust in institutions and electoral processes,[170] and in conflict they can increase escalatory risks, such as by complicating military campaigns at the tactical level or increasing fears of pre-emptive attacks. During a nuclear crisis, for example, a State may attempt to influence the domestic debates of an adversary

[165] Boulanin et al., "Artificial Intelligence, Strategic Stability and Nuclear Risk", 25–26; Zachary Kallenborn, "AI Risks to Nuclear Deterrence are Real", *War on the Rocks*, 10 October 2019, https://warontherocks.com/2019/10/ai-risks-to-nuclear-deterrence-are-real/.

[166] Boulanin et al., "Artificial Intelligence, Strategic Stability and Nuclear Risk", 108–109.

[167] Johnson, "Inadvertent Escalation in the Age of Intelligent Machines", 349.

[168] Ibid., 349; Boulanin et al., "Artificial Intelligence, Strategic Stability and Nuclear Risk", 27.

[169] Interview Andrew Lohn (16 February 2023).

[170] Matteo E. Bonfanti, "The Weaponisation of Synthetic Media: What threat does this pose to national security?" Elcano Royal Institute, 14 July 2020, https://www.realinstitutoelcano.org/en/analyses/the-weaponisation-of-synthetic-media-what-threat-does-this-pose-to-national-security/.

in order to exert pressure on the leadership.[171]

Further, the impact of AI on disinformation can be reflected insidiously in military intelligence, with effects that can be far-reaching and difficult to contain. For example, as open-source intelligence relies massively on public sources, the contamination of public information with widespread fake and misleading data can reflect back into the kind of outputs delivered to military intelligence.

Incident Scenario

TYPE OF INCIDENT AND CONTEXT: AUTOMATIC MISSILE LAUNCH	POSSIBLE ESCALATORY CONSEQUENCES
An air defence system used to provide protection against incoming missiles around a military base uses AI to process gigabytes of data in real time. An unusual glare on the horizon one day is misclassified by the algorithm as a missile attack. The system is equipped to respond in eight seconds from the moment it detected the threat and subsequently starts to launch interceptors.[172]	At the other end, other States may order retaliatory counter-attacks, as their early-warning systems respond to what is now the incoming threat (i.e., interceptor launch/nuclear weapons).
An extreme version of this scenario would involve **autonomous nuclear launch platforms:** An autonomous launch platform designed to retaliate against incoming nuclear weapons concludes a nuclear offensive is about to begin and launches a nuclear weapon. The signal it responded to was coming from a test that another State was conducting and which involved the launch of rockets into the atmosphere. Data poisoning interfered with the system's algorithms and changed the parameters of classifiers in the system.[173]	

[171] Johnson, "Inadvertent Escalation in the Age of Intelligent Machines", 346–347.

[172] A similar risk scenario was introduced in Arnold and Toner, "AI Accidents". For a review of relevant technologies, see Vincent Boulanin and Maaike Verbruggen, "Mapping the Development of Autonomy in Weapon Systems", Stockholm International Peace Research Institute, November 2017, 36–39. https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf.

[173] For a discussion of similar scenarios/use cases see, for example, Horowitz and Scharre, "AI and International Stability,"; Zachary Kallenborn, "Giving an AI Control of Nuclear Weapons: What could possibly go wrong?" Bulletin of the Atomic Scientists, 1 February 2022, https://thebulletin.org/2022/02/giving-an-ai-control-of-nuclear-weapons-what-could-possibly-go-wrong/.

3. Risks of Proliferation

AI can enable the proliferation of new weapons, such as through convergence with other fields of science and technology, or by enhancing the lethality and autonomy of existing weapon systems, and thus the appeal for their proliferation and acquisition.

Convergence Risks

a. Biosecurity and Chemical Weapons

The convergence between AI, biology and chemistry has created opportunities for advances in the medical field and drug discovery yet it is also a perfect illustration of dual-use risks. For example, a class of LLMs called chemical language models (CLMs) are being employed to discover new therapies and to predict, among others, potential drug molecules that target specific proteins which cause diseases.[174] AI language models can be applied to generate new proteins (e.g., ProtGPT2), and while potentially advancing solutions to fight disease, such applications can create opportunities for misuse.[175] In recent years, research and policy communities have turned their attention to the emerging risks of misuse of AI in the domain of biotechnology. One critical risk concerns the proliferation of biochemical weapons, though risks are more diverse and with multiple layers of complexity.

For a start, the convergence between AI and biotechnology may lower the tacit knowledge that was required for tedious laboratory tasks, or render complex notions and concepts more easily understandable. This means that many more actors can have access to the life sciences.[176] This also renders the scope of risks very broad: from uses of AI to assist in the identification of virulence factors to *in silico* design of new pathogens.[177]

Research published in 2022 has showed that commercial and open-source ML software can be employed for de novo design processes of new molecules, many of which were shown to be toxic. From computational proof in a lab setting to physical synthesis of new molecules, the barriers are small, particularly in the context of a surge in commercial companies worldwide which offer chemical synthesis.[178]

[174] Steph Batalis, Caroline Schuerger, and Vikram Venkatram, "Large Language Models in Biology", Center for Security and Emerging Technology, 16 June 2023, https://cset.georgetown.edu/article/large-language-models-in-biology/.

[175] Sean Ekins et al., "There's a 'ChatGPT' for Biology. What could go wrong?" *Bulletin of the Atomic Scientists*, 24 March 2023, https://thebulletin.org/2023/03/chat-gpt-for-biology/.

[176] John T. O'Brien and Cassidy Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology", *Health Security*, Vol. 18, No. 3 (May/June 2020), 220, https://www.liebertpub.com/doi/epdf/10.1089/hs.2019.0122; interview with Dan Hendrycks (27 April 2023).

[177] O'Brien and Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology", 220.

[178] Fabio Urbina et al., "Dual Use of Artificial-Intelligence-Powered Drug Discovery", *Nature Machine Intelligence*, Vol. 4 (2022), 189–190, https://doi.org/10.1038/s42256-022-00465-9.

To add further complexity to the risks of convergence of AI and biotechnology, additional vulnerabilities can be caused by offensive cyberattacks. These could target vulnerabilities in DNA synthesizers (which are machines that build custom-made DNA molecules) and introduce malware that could disrupt the DNA sequences, or target the biosecurity of the laboratory.[179]

b. Cybersecurity

AI introduces additional proliferation risks in the cyber domain.

The potential use of AI to develop malicious code has been a key concern in this area of convergence, before recent breakthroughs in LLMs.[180] LLMs have exposed additional risks, including risks of LLMs being requested to generate malware. Research on adversarial attacks (discussed in Part I), has shown that large language models may not always be effective at detecting inputs from malicious actors, or inputs that are disguised as benign.

Large language models, which include LLMs that are widely used for coding,[181] benefit from thorough red teaming by cybersecurity experts, who assess vulnerability to cyberattacks. Such systems are typically trained and tested for their ability to refuse to write malware, but the risk cannot be ruled out, especially as **open source models** are proliferating.

Further, risks in the field of cybersecurity are not only about ease of execution but also about scope and scale. While various forms of subversion and attack are possible with existing cyber offence tools, advances in AI can supercharge cybersecurity threats, including by enhancing the speed, power and scale of attacks in cyberspace.[182] The integration of ML in a growing number of technical systems, ranging from satellite navigation to intelligence, surveillance and reconnaissance missions, means that attacks could be executed at scale and across domains.[183]

Proliferation of AI and Autonomous Weapons Systems

a. Proliferation of AI

In addition to more specific concerns about the proliferation of AI-enabled autonomous weapons, discussed below, the digital nature of

[179] O'Brien and Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology," 223; Sina Faezi et al., "Oligo-Snoop: A non-invasive side channel attack against DNA synthesis machines", Network and Distributed Systems Security (NDSS) Symposium 2019, 24–27 February 2019, San Diego, USA, https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_05B-1_Faezi_paper.pdf.

[180] Emilia Javorsky and Hamza Chaudhry, "Convergence: Artificial intelligence and the new and old weapons of mass destruction", *Bulletin of the Atomic Scientists*, 18 August 2023, https://thebulletin.org/2023/08/convergence-artificial-intelligence-and-the-new-and-old-weapons-of-mass-destruction/.

[181] For example, Meta's Code Llama, released in August 2023.

[182] James Johnson, "The AI-Cyber Nexus: Implications for military escalation, deterrence and strategic stability", *Journal of Cyber Policy*, Vol. 4, No. 3 (2019), 449, https://doi.org/10.1080/23738871.2019.1701693.

[183] Ibid.; Although not a proliferation risk, the cybersecurity risks of AI systems, discussed earlier in the report, open the possibility for attackers to gain access to the machine learning algorithms or training datasets, which can effectively offer access to vast amounts of data used in defence systems.

AI means the integration and use of the technology can diffuse rapidly across domains of use and across borders. In other words, risks of proliferation in the context of AI must first factor in the proliferation of the technology itself.

AI-powered software, once developed, can be repurposed at minimal cost, which makes it difficult in practice to prevent State or non-State actors, including those with hostile intent, from adopting the technology.[184] As knowledge, skills and resource barriers are lowered, the range of risks rapidly increases.

Both irresponsible State actors and malicious non-State actors could integrate AI into a range of conventional weapons systems (for example, weapons which otherwise lack autonomous functions), and efforts to conduct complex cyberattacks at scale may be boosted by new capabilities afforded by LLMs.

b. Proliferation of Autonomous Weapons

The proliferation of (lethal) autonomous weapons systems[185] has been a growing concern for the international community. Already, the removal of humans from the battlefield in some contexts, combined with the capabilities of autonomous systems, have rendered the pace of performance beyond human ability for intervention. Contemporary systems that can autonomously calculate flight paths for rockets and the location for intercepting missiles are faster than the ability of any human, no matter how skilled, to operate.[186] While these operational assets can provide incentives for many States, they entail new types of risks, including risks for the management of escalation (highlighted above).

However, there are increasing concerns about the proliferation of autonomous weapons in light of a **growing use of uncrewed systems**. Unlike complex future systems, uncrewed systems incur relatively low costs of development and deployment, which is also due to the fact that they do not necessarily have to be fitted with defensive systems.[187] Further, achieving a relatively simple level of autonomy is not very difficult.[188]

At a higher end, however, there are systems developed by States or private manufactures

[184] Edward Hunter Christie and Amy Ertan, “NATO and Artificial Intelligence”, 20 December 2021, Routledge Companion to Artificial Intelligence and National Security Policy, forthcoming, (eds.) Romaniuk, S. N., and Manjikian. M. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4133397.

Fast and virtually unrestricted diffusion of AI technology complicates existing international efforts in the area of non-proliferation and export controls.

[185] The definition of autonomous weapons systems (AWS) remains contested, and an extensive discussion is beyond the scope of this report. The working definition applied here is the definition commonly used in policy debates, which define AWS as weapons systems that, once activated, can engage targets without further human intervention.

[186] Liran Antebi, “The Proliferation of Autonomous Weapons Systems: Effects on International Relations”, in National Security in a “Liquid” World, Eds. Carmit Padan and Vera Michlin-Shapir, (Tel Aviv: The Institute for National Security Studies, Memorandum 195, October 2019), 84–85, https://www.inss.org.il/wp-content/uploads/2019/10/Memo195_e_compressed.pdf.

[187] Antebi, “The Proliferation of Autonomous Weapons Systems”, 85.

[188] Zachary Kallenborn, “Applying Arms-Control Frameworks to Autonomous Weapons”, Brookings Institution, 5 October 2021, https://www.brookings.edu/articles/applying-arms-control-frameworks-to-autonomous-weapons/. Advances with loitering munitions have known steady progress and a real, though still relatively limited, impact in military operations.

with significantly advanced autonomous functions, including potential for operation without a human in the loop. Loitering munitions produced and tested in recent years demonstrate an expanding array of autonomous functions. These include systems that can hover, locate, and track both stationary and moving targets, with many operating as part of larger (integrated) systems. This means that, for example, the targeting information can be leveraged from multiple sources, including cameras fitted to the system, as well as data retrieved from other surveillance drones.[189]

Further, the development of lethal payloads for uncrewed systems is a further incentive for proliferation, and advances in swarm robotics, powered by AI, can increase their capacity to be disruptive and devastating in conflict, including by enabling distributed attacks or saturating the defences of an adversary.[190]

These capabilities have been bolstered by steady advances in other technical domains (e.g., material science or energy), which afford greater operational endurance, including heavier payloads and longer deployment periods.

Against the backdrop of continuous scaling of autonomous functions, there are already many uncertainties as to how these systems are operated, and perhaps even more critically, where the tipping point might be for relinquishing the capacity for human intervention in favour of full autonomy (i.e., including over target selection and engagement). The fact that certain autonomous functions remain latent for now may shift and evolve in the context of rapidly evolving circumstances in armed conflict, and as competitive pressures mount to deploy faster and more lethal systems.

[189] A catalogue created by AutoNorms, an international research project hosted at the University of Southern Denmark, lists key technical factsheets for a global sample of 24 loitering munitions. The comparative analysis of fielded systems shows clear trends for continuous development of autonomous functions; see Tom Watts and Ingvild Bode, “Automation and Autonomy in Loitering Munitions Catalogue (v.1)”, 25 April 2023, **https://doi.org/10.5281/zenodo.7860762**.

[190] For a comprehensive overview of technological developments in the area of uncrewed systems, see Sarah Grand-Clement, “Uncrewed Aerial, Ground, and Maritime Systems: A Compendium”, UNIDIR, 3 April 2023, **https://unidir.org/publication/uncrewed-aerial-ground-and-maritime-systems-compendium**.

Conclusion

Some of the worst-case scenarios involving AI technology have fortunately not been verified to date, and so they remain to a certain extent in the ambit of speculation. This speculation warrants a certain degree of caution when discussing AI's escalatory potential in conflict, or dangers of proliferation. However, a survey of what is possible with the state of the technology today provides sufficient grounds to forecast the technology's transformative potential for global security. These concerns must be heeded carefully. **Risk mitigation begins with a comprehensive and informed understanding of risks.** The next step is a collaborative effort to work for a more transparent and safe use of artificial intelligence in the military domain and in the context of international security, broadly.

This research report presented a taxonomy of risks of AI in the context of international peace and security, divided in two large clusters of risks: **risks of the technology** (safety, security, and human–machine interaction risks) and **risks of AI to global security** (miscalculation, escalation, and proliferation). These risks are closely interrelated, and though they can be taxonomized in different categories, they may ultimately require an inclusive approach to risk management. Further, unlike other capabilities or domains, AI is not a discrete but a general-purpose technology, which means that the elaboration of CBMs may require different or more innovative approaches. Lessons from the cyber domain may prove useful but not be entirely transferable.

There are many options that lie ahead for a future elaboration of confidence-building measures for AI. These options are explored in the next phase of this project, with participation and co-ownership from diverse stakeholders.

Potential pathways for future CBMs may consider **questions** such as:

- What actionable steps can States take to elaborate CBMs at the multilateral level?
- (How) should national and regional efforts contribute to the elaboration of CBMs?
- What is the best forum or framework for States to launch a dedicated process for developing CBMs for AI?
- What are the most realistic and feasible approaches to develop CBMs?
- What instruments (e.g., verification mechanisms) should support CBMs?
- Which actors need to be involved in the development of CBMs?

AI can bring many opportunities and benefits in the context of international security, and the areas of risks discussed in this report are a flip side of the technology's full potential. These risks are, however, real and they must be understood and addressed to harness the technology's extraordinary capabilities.

CBMs can help chart a more transparent, safe and responsible development and/or deployment of the technology, without any prejudice to future arms control regimes or any legally binding instruments that may emerge. UNIDIR will support multilateral efforts in that direction.

Bibliography

Adhikari, Ajaya, Richard den Hollander, Ioannis Tolios, Michael van Bekkum, Anneloes Bal, Stijn Hendriks, Maarten Kruithof, Dennis Gross, Nils Jansen, Guillermo Pérez, Kit Buurman, and Stephan Raaijmakers. "Adversarial Patch Camouflage Against Aerial Detection." arXiv (31 August 2020): 1-9. https://arxiv.org/pdf/2008.13671.pdf.

Agahari, Wirawan, Hosea Ofe, and Mark de Reuver. "It is not (only) about privacy: How multiparty computation redefines control, trust, and risk in data sharing." Electronic Markets 32 (September 2022): 1577-1602. https://doi.org/10.1007/s12525-022-00572-w.

Agrawal, Pulkit. "The Task Specification Problem." Proceedings of the 5th Conference on Robot Learning. Proceedings of Machine Learning Research Vol. 164 (2022): 1-7. https://proceedings.mlr.press/v164/agrawal22a/agrawal22a.pdf.

Amer K., M. Samy, M. Shaker, and M. ElHelw. "Deep Convolutional Neural Network-Based Autonomous Drone Navigation." arXiv (5 May 2019). https://arxiv.org/abs/1905.01657.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety." arXiv (25 July 2016): 1-29. https://arxiv.org/pdf/1606.06565.pdf.

Antebi, Liran. "The Proliferation of Autonomous Weapons Systems: Effects on International Relations." In National Security in a "Liquid" World, edited by Carmit Padan and Vera Michlin-Shapir, 75-92. Tel Aviv: The Institute for National Security Studies, Memorandum 195, October 2019. https://www.inss.org.il/wp-content/uploads/2019/10/Memo195_e_compressed.pdf.

Arnold, Zachary and Helen Toner. "AI Accidents: An Emerging Threat. What Could Happen and What to Do." Center for Security and Emerging Technology. July 2021. https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/.

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. "Synthesizing Robust Adversarial Examples." arXiv, 7 June 2018. Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. https://arxiv.org/pdf/1707.07397.pdf.

Bagdasaryan, Eugene, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. "(Ab)using Images and Sounds for Indirect Instruction Injection in Multi-Modal LLMs." arXiv (24 July 2023). https://arxiv.org/abs/2307.10490.

Bainbridge, Lisanne. "Ironies of Automation." Automatica Vol. 19, 6 (November 1983): 775-779. https://doi.org/10.1016/0005-1098(83)90046-8.

Bajema, Natasha. "Pentagon Wants AI to Predict Events Before They Occur." IEEE Spectrum (14 October 2021). https://spectrum.ieee.org/predictive-ai-pentagon.

Barnes, Michael J. and A. William Evans III. "Soldier-Robot Teams in Future Battlefields: An Overview." In Human-Robot Interactions in Future Military Operations, edited by Michael Barnes and Florian Jentsch, 9- 29. Boca Raton: CRC Press, 2017.

Batalis, Steph, Caroline Schuerger, and Vikram Venkatram. "Large Language Models in Biology." Center for Security and Emerging Technology. 16 June 2023. https://cset.georgetown.edu/article/large-language-models-in-biology/.

Bobu, Andreea, Andi Peng, Pulkit Agrawal, Julie Shah, and Anca D. Dragan. "Aligning Robot and Human Representations." arXiv (3 February 2023). https://arxiv.org/abs/2302.01928.

Bonfanti, Matteo E. "The weaponisation of synthetic media: what threat does this pose to national security?" Elcano Royal Institute. 14 July 2020. https://www.realinstitutoelcano.org/en/analyses/the-weaponisation-of-synthetic-media-what-threat-does-this-pose-to-national-security/.

Boulanin, Vincent, Lora Saalman, Petr Topychkanov, Fei Su, and Moa Peldán Carlsson. "Artificial Intelligence, Strategic Stability and Nuclear Risk." Stockholm International Peace Research Institute. June 2020. https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

Boulanin, Vincent and Maaike Verbruggen. "Mapping the Development of Autonomy in Weapon Systems." Stockholm International Peace Research Institute. November 2017. https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf.

C, Adam and Richard Carter. "Large Language Models and Intelligence Analysis." Centre for Emerging Technology and Security (CETaS Expert Analysis). July 2023. https://cetas.turing.ac.uk/sites/default/files/2023-07/cetas_expert_analysis_-_large_language_models_and_intelligence_analysis.pdf.

Carlini, Nicholas, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, Florian Tramèr. “Membership Inference Attacks From First Principles.” arXiv (12 April 2022): 1-20. https://arxiv.org/pdf/2112.03570.pdf.

Christie, Edward Hunter and Amy Ertan. “NATO and Artificial Intelligence.” (20 December 2021). In Routledge Companion to Artificial Intelligence and National Security Policy, forthcoming, edited by S. N. Romaniuk and Manjikian. M. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4133397.

Cuevas, Haydee M., Stephen M Fiore, Barrett S Caldwell, and Laura Strater. “Augmenting Team Cognition in Human-Automation Teams Performing in Complex Operational Environments.” Aviation, Space and Environmental Medicine 78:5 Section II (May 2007): B63-70. https://pubmed.ncbi.nlm.nih.gov/17547306/.

Cummings, Mary L. “Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings.” AI Magazine 42, No. 1 (Spring 2021): 6-15. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/7394.

Cummings, Mary L. and Songpo Li. “Subjectivity in the Creation of Machine Learning Models.” Journal of Data and Information Quality 13, Issue 2, Article 7 (May 2021): 1-19. https://doi.org/10.1145/3418034.

Cummings, Mary L. “Automation and Accountability in Decision Support System Interface Design.” The Journal of Technology Studies Vol. 32, Number 1 (Winter 2006): 23-31. https://jotsjournal.org/articles/10.21061/jots.v32i1.a.4 .

Cummings, Mary. “Automation Bias in Intelligent Time Critical Decision Support Systems.” AIAA 1st Intelligent Systems Technical Conference (Chicago, Illinois). American Institute of Aeronautics and Astronautics. 20-22 September 2004 (Published online: 19 June 2012). https://doi.org/10.2514/6.2004-6313.

Cummings, M.L. “Revisiting human-systems engineering principles for embedded AI applications.” Frontiers in Neuroergonomics Vol. 4 (January 2023). https://doi.org/10.3389/fnrgo.2023.1102165.

Cummings, Missy. “Identifying AI Hazards and Responsibility Gaps” (draft). July 2023. https://www.researchgate.net/publication/372051108_Identifying_AI_Hazards_and_Responsibility_Gaps.

Dai, Yan, Dan Liu, Qingrong Hu, and Xiaoli Yu. “Radar Target Detection Algorithm Using Convolutional Neural Network to Process Graphically Expressed Range Time Series Signals.” Sensors 22, No 18. 6868 (11 September 2022). https://doi.org/10.3390/s22186868.

de Bruijn, Hans, Martijn Warnier, and Marijn Janssen. “The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making.” Government Information Quarterly Vol. 39, Issue 2 (April 2022): 1-8. https://doi.org/10.1016/j.giq.2021.101666.

Delcker, Janosch. “Meteorologists of violence.” Politico (15 March 2020). https://www.politico.eu/article/artificial-intelligence-conflict-war-prediction/.

Development, Concepts and Doctrine Center (Ministry of Defence, United Kingdom). ”Human-Machine Teaming.” Joint Concept Note 1/18. May 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/709359/20180517-concepts_uk_human_machine_teaming_jcn_1_18.pdf.

DoD Responsible AI Working Council. U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway. June 2022. https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf.

Eckersley, Peter. “The Cautious Path to Strategic Advantage: How Militaries Should Plan for AI.” Electronic Frontier Foundation. 2018. https://www.eff.org/files/2018/10/12/the_cautious_path_to_strategic_advantage_how_militaries_should_plan_for_ai_v1.1_0.pdf.

Ekelhof, Merel and Giacomo Persi Paoli. “The Human Element in Decisions about the Use of Force.” UNIDIR. 2019. https://unidir.org/sites/default/files/2020-03/UNIDIR_Iceberg_SinglePages_web.pdf.

Ekins, Sean, Filippa Lentzos, Max Brackmann, and Cédric Invernizzi. “There’s a ‘ChatGPT’ for biology. What could go wrong?” Bulletin of the Atomic Scientists (24 March 2023). https://thebulletin.org/2023/03/chat-gpt-for-biology/.

Endsley, Mica R. “Supporting Human-AI Teams: Transparency, explainability and situation awareness.” Computers in Human Behavior Vol. 140 (March 2023). https://doi.org/10.1016/j.chb.2022.107574.

Endsley, Mica R. “Autonomous Driving Systems: A Preliminary Naturalistic Study of the Tesla Model S.” Journal of Cognitive Engineering and Decision Making Vol. 11, Issue 3 (2017): 225-238. https://doi.org/10.1177/1555343417695197.

Endsley, Mica R. “Toward a Theory of Situation Awareness in Dynamic Systems.” Human Factors: The Journal of the Human Factors and Ergonomics Society Vol. 37, Issue 1 (March 1995): 32-64. https://doi.org/10.1518/001872095779049543.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. “Robust Physical-World Attacks on Deep Learning Visual Classification.” arXiv (10 April 2018). https://arxiv.org/pdf/1707.08945.pdf.

Faezi, Sina, Sujit Rokka Chhetri, Arnav Vaibhav Malawade, John Charles Chaput, William Grover, Philip Brisk, and Mohammad Abdullah Al Faruque. “Oligo-Snoop: A Non-Invasive Side Channel Attack Against DNA Synthesis Machines.” Network and Distributed Systems Security (NDSS) Symposium 2019. 24-27 February 2019. San Diego, USA. **https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_05B-1_Faezi_paper.pdf**.

Falco, Gregory and Nicolo Boschetti. “A Security Risk Taxonomy for Commercial Space Missions”. ASCEND. 15-17 November 2021, Las Vegas and Virtual. **https://doi.org/10.2514/6.2021-4241**.

Federal Office for Information Security (Germany). “AI Security Concerns in a Nutshell.” 9 March 2023. **https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_Al-Security_Guide_2023.pdf?__blob=publicationFile&v=5**.

Flemisch, Frank O., Eugen Altendorf, Yigiterkut Canpolat, Gina Weßel, Marcel Baltzer, Daniel Lopez, Nicolas Daniel Herzberger, Gudrun Mechthild Irmgard Voß, Maximilian Schwalm, and Paul Schutte. “Uncanny and Unsafe Valley of Assistance and Automation: First Sketch and Application to Vehicle Automation.” Advances in Ergonomic Design of Systems, Products and Processes: Proceedings of the Annual Meeting of the GfA 2016 (Springer: Berlin & Heidelberg, 2017): 319-334. **https://doi.org/10.1007/978-3-662-53305-5_23**.

Flournoy, Michèle A., Avril Haines, and Gabrielle Chefitz. “Building Trust through Testing. Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems.” Center for Security and Emerging Technology. October 2020. **https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf**.

Ghioni, Riccardo, Mariarosaria Taddeo, and Luciano Floridi. “Open source intelligence and AI: a systemic review of the GELSI literature.” AI & Society (2023). **https://doi.org/10.1007/s00146-023-01628-x**.

Goodfellow, Ian and Nicolas Papernot. “Is attacking machine learning easier than defending it?” Cleverhans Blog (15 February 2017). **http://www.cleverhans.io/security/privacy/ml/2017/02/15/why-attacking-machine-learning-is-easier-than-defending-it.html**.

Grand-Clement, Sarah. “Uncrewed Aerial, Ground, and Maritime Systems: A Compendium.” UNIDIR. 3 April 2023. **https://unidir.org/publication/uncrewed-aerial-ground-and-maritime-systems-compendium**.

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.

---------- “Report of the 2023 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.” CCW/GGE.1/2023/2. 24 May 2023. **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_2_Advance_version.pdf**.

---------- Draft Protocol on Autonomous Weapon Systems (Protocol VI), Submitted by Argentina, Ecuador, El Salvador, Colombia, Costa Rica, Guatemala, Kazakhstan, Nigeria, Palestine, Panama, Peru, Philippines, Sierra Leone and Uruguay. 11 May 2023. CCW/GGE.1/2023/WP.6. **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.6_2.pdf**.

----------- Proposal for an international legal instrument on Lethal Autonomous Weapons Systems (LAWS), Submitted by Pakistan. 8 March 2023. CCW/GGE.1/2023/WP.3/Rev.1. **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.3_REv.1_0.pdf**.

----------- Concept of Activities of the Armed Forces of the Russian Federation in the Development and Use of Weapons Systems with Artificial Intelligence Technologies (unofficial translation), Submitted by the Russian Federation. 7 March 2023. CCW/GGE.1/2023/WP.5. **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.5_0.pdf**.

---------- State of Palestine’s Proposal for the Normative and Operational Framework on Autonomous Weapons Systems, submitted by the State of Palestine. 3 March 2023. CCW/GGE.1/2023/WP.2/Rev1. **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.2_Rev.1.pdf**.

Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain.” arXiv (11 March 2019). **https://arxiv.org/pdf/1708.06733.pdf**.

Hawley, John K. “Patriot Wars. Automation and the Patriot Air and Missile Defense System.” Center for a New American Security. January 2017. **https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Report-EthicalAutonomy5-PatriotWars-FINAL.pdf**.

Hoff, Kevin Anthony and Masooda Bashir. “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust.” Human Factors: The Journal of the Human Factors and Ergonomics Society Vol. 57, Issue 3 (May 2015): 407-434. https://doi.org/10.1177/0018720814547570.

Hoffman, Wyatt. “AI and the Future of Cyber Competition.” Center for Security and Emerging Technology. January 2021. https://cset.georgetown.edu/publication/ai-and-the-future-of-cyber-competition/.

Hoffman, Wyatt and Heeu Millie Kim. “Reducing the Risks of Artificial Intelligence for Military Decision Advantage.” Center for Security and Emerging Technology. March 2023. https://cset.georgetown.edu/publication/reducing-the-risks-of-artificial-intelligence-for-military-decision-advantage/.

Holland Michel, Arthur. “The Black Box, Unlocked. Predictability and Understandability in Military AI.” UNIDIR. 22 September 2020. https://unidir.org/sites/default/files/2020-09/BlackBoxUnlocked.pdf.

Horowitz, Michael C. and Lauren Kahn. “Bending the Automation Bias Curve: A study of human and AI-based decision making in national security contexts.” arXiv (30 June 2023). https://arxiv.org/abs/2306.16507.

Horowitz, Michael C. and Lauren Kahn. “Leading in Artificial Intelligence through Confidence Building Measures.” The Washington Quarterly 44, Issue 4 (Winter 2021): 91-106. https://doi.org/10.1080/0163660X.2021.2018794.

Horowitz, Michael and Paul Scharre. “AI and International Stability: Risks and Confidence-Building Measures.” Center for a New American Security. January 2021. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

IBM. “What is Federated Learning.” 24 August 2022. https://research.ibm.com/blog/what-is-federated-learning.

Ilyas, Andrew, Logan Engstrom, Anish Athalye, and Jessy Lin. “Black-box Adversarial Attacks with Limited Queries and Information.” arXiv (11 July 2018). https://arxiv.org/abs/1804.08598.

International Committee of the Red Cross. International Humanitarian Law Databases. Rule 57, “Ruses of War.” https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_cha_chapter18_rule57.

International Organization for Standardization. 31000:2018. “Risk management guidelines,” 2018. https://www.iso.org/obp/ui/en/#iso:std:iso:31000:ed-2:v1:en.

International Organization for Standardization. ISO 9241-11:2018. “Ergonomics of human-system interaction – Part 11: Usability: definitions and concepts.” https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

Irolla, Paul. “What is model stealing and why it matters.” ML Security (23 December 2019). https://www.mlsecurity.ai/post/what-is-model-stealing-and-why-it-matters.

Janjeva, Ardi, Nikhil Mulani, Rosamund Powell, Jess Whittlestone, and Shahar Avin. “Strengthening Resilience to AI Risk. A guide for UK policymakers.” Centre for Emerging Technology and Security & Centre for Long-Term Resilience. August 2023. https://cetas.turing.ac.uk/publications/strengthening-resilience-ai-risk.

Javorsky, Emilia and Hamza Chaudhry. “Convergence: Artificial intelligence and the new and old weapons of mass destruction.” Bulletin of the Atomic Scientists (18 August 2023). https://thebulletin.org/2023/08/convergence-artificial-intelligence-and-the-new-and-old-weapons-of-mass-destruction/.

Jensen, Benjamin and Dan Tadross. “How Large-Language Models Can Revolutionize Military Planning.” War on the Rocks (12 April 2023). https://warontherocks.com/2023/04/how-large-language-models-can-revolutionize-military-planning/.

Johnson, James. “Inadvertent escalation in the age of intelligent machines: A new model for nuclear risk in the digital era.” European Journal of International Security 7, Issue 3 (2022): 337-359. https://doi.org/10.1017/eis.2021.23.

Johnson, James. “The AI-cyber nexus: implications for military escalation, deterrence and strategic stability.” Journal of Cyber Policy 4, Number 3 (2019): 442-460. https://doi.org/10.1080/23738871.2019.1701693.

Kahn, Lauren. “Mending the “Broken Arrow”: Confidence Building Measures at the AI-Nuclear Nexus.” War on the Rocks (4 November 2022). https://warontherocks.com/2022/11/mending-the-broken-arrow-confidence-building-measures-at-the-ai-nuclear-nexus/.

Kallenborn, Zachary. “Giving an AI control of nuclear weapons: What could possibly go wrong?” Bulletin of the Atomic Scientists (1 February 2022). https://thebulletin.org/2022/02/giving-an-ai-control-of-nuclear-weapons-what-could-possibly-go-wrong/.

Kallenborn, Zachary. “AI Risks to Nuclear Deterrence are Real.” War on the Rocks (10 October 2019). https://warontherocks.com/2019/10/ai-risks-to-nuclear-deterrence-are-real/.

Kallenborn, Zachary. “Applying arms-control frameworks to autonomous weapons.” Brookings Institution. 5 October 2021. https://www.brookings.edu/articles/applying-arms-control-frameworks-to-autonomous-weapons/.

Kiran, B Ravi, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. “Deep Reinforcement Learning for Autonomous Driving: A Survey.” arXiv (23 January 2021): 1-18. https://arxiv.org/pdf/2002.00444.pdf.

Knack, Anna, Richard J. Carter, and Alexander Babuta. “Human-Machine Teaming in Intelligence Analysis. Requirements for developing trust in machine learning systems.” Centre for Emerging Technology and Security. December 2022. https://cetas.turing.ac.uk/sites/default/files/2022-12/cetas_research_report_-_hmt_and_intelligence_analysis_vfinal.pdf.

Konaev, Margarita, Tina Huang, and Husanjot Chahal. “Trusted Partners. Human-Machine Teaming and the Future of Military AI.” Center for Security and Emerging Technology. February 2021. https://cset.georgetown.edu/publication/trusted-partners/.

Lee, John D. and Katrina A. See. “Trust in Automation: Designing for Appropriate Reliance.” Human Factors: The Journal of the Human Factors and Ergonomics Society Vol. 46, Issue 1 (Spring 2004): 50-80. https://journals.sagepub.com/doi/epdf/10.1518/hfes.46.1.50_30392.

Lin-Greenberg, Erik. “Evaluating Escalation: Conceptualizing Escalation in an Era of Emerging Military Technologies.” The Journal of Politics 85, Number 3 (July 2023): 1151-1155. https://www.journals.uchicago.edu/doi/full/10.1086/723974.

Lohn, Andrew J. “Hacking AI. A Primer for Policymakers on Machine Learning Cybersecurity.” Center for Security and Emerging Technology. December 2020. https://cset.georgetown.edu/publication/hacking-ai/.

Lohn, Andrew J. “Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-of-Distribution Performance.” arXiv (2 September 2020): 1-15. https://arxiv.org/pdf/2009.00802.pdf.

Mecklin, John. “Interview: Emerging military technology expert Paul Scharre on global power dynamics in the AI age.” Bulletin of the Atomic Scientists (11 September 2023). https://thebulletin.org/premium/2023-09/interview-emerging-military-technology-expert-paul-scharre-on-global-power-dynamics-in-the-ai-age/.

Morgan, Forrest E., Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima and Derek Grossman. “Military Applications of Artificial Intelligence. Ethical Concerns in an Uncertain World.” RAND. 2020. https://www.rand.org/pubs/research_reports/RR3139-1.html.

Musser, Micah, Andrew Lohn, James X. Dempsey, Jonathan Spring, Ram Shankar Siva Kumar, Brenda Leong, Christina Liaghati, Cindy Martinez, Crystal D. Grant, Daniel Rohrer, Heather Frase, John Bansemer, Jonathan Elliott, Mikel Rodriguez, Mitt Regan, Rumman Chowdhury, and Stefan Hermanek. “Adversarial Machine Learning and Cybersecurity. Risks, Challenges, and Legal Implications.” Center for Security and Emerging Technology. April 2023. https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/.

National Academies of Sciences, Engineering, and Medicine. Human-AI Teaming: State-of-the-Art and Research Needs. Washington DC: The National Academies Press, 2022. https://doi.org/10.17226/26355.

O’Brien, John T. and Cassidy Nelson. “Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology.” Health Security 18, Issue 3 (May/June 2020): 219-227. https://www.liebertpub.com/doi/epdf/10.1089/hs.2019.0122.

O’Neill, Thomas, Nathan McNeese, Amy Barron, and Beau Schelble. “Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature.” Human Factors: The Journal of the Human Factors and Ergonomics Society Vol. 64, Issue 5 (August 2022):904-938. https://doi.org/10.1177/0018720820960865.

Organisation for Economic Cooperation and Development. “Advancing Accountability in AI. Governing and managing risks throughout the lifecycle for trustworthy AI.” OECD Digital Economy Papers. No 349. 23 February 2023. https://doi.org/10.1787/2448f04b-en.

OpenAI. “Attacking machine learning with adversarial examples.” 24 February 2017. https://openai.com/research/attacking-machine-learning-with-adversarial-examples.

Palantir. “Artificial Intelligence Platform for Defense.” https://www.palantir.com/platforms/aip/.

Papernot, Nicolas, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. “Practical Black-Box Attacks against Machine Learning.” ASIA CCS ’17: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (April 2017). https://doi.org/10.1145/3052973.3053009.

Parasuraman, Raja, Michael Barnes, and Keryl Cosenzo. “Adaptive Automation for Human-Robot Teaming in Future Command and Control Systems.” The International C2 Journal Vol. 1, Number 2 (2007): 43-68. https://apps.dtic.mil/sti/pdfs/ADA503770.pdf.

Puscas, Ioana. “Confidence-Building Measures for Artificial Intelligence: A Framing Paper.” UNIDIR. 19 December 2022. https://unidir.org/publication/confidence-building-measures-artificial-intelligence-framing-paper.

Puscas, Ioana. “Human-Machine Interfaces in Autonomous Weapons Systems.” UNIDIR. 21 July 2022. https://www.unidir.org/publication/human-machine-interfaces-autonomous-weapon-systems.

Ratches, James A. “Review of current aided/automatic target acquisition technology for military target acquisition tasks.” Optical Engineering Vol. 50, Issue 7 (July 2011): 072001-1 – 072001-8. https://doi.org/10.1117/1.3601879.

Reynolds, Ian and Ozan Ahmet Cetin. “War is messy. AI can’t handle it.” Bulletin of the Atomic Scientists (14 August 2023). https://thebulletin.org/2023/08/war-is-messy-ai-cant-handle-it/#post-heading.

Riley, Jennifer M., Laura D. Strater, Sheryl L. Chappell, Erik S. Connors, and Mica R. Endsley. “Situation Awareness in Human-Robot Interaction: Challenges and User Interface Requirements.” In Human-Robot Interactions in Future Military Operations, edited by Michael Barnes and Florian Jentsch, 171-191. Boca Raton: CRC Press, 2017.

Rudner, Tim G.J. and Helen Toner. “Key Concepts in AI Safety: Specification in Machine Learning.” Center for Security and Emerging Technology. December 2021. https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-specification-in-machine-learning/.

Rudner, Tim G.J. and Helen Toner. “Key Concepts in AI Safety: Robustness and Adversarial Examples.” Centre for Security and Emerging Technology. March 2021. https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/.

Russell, Stuart and Peter Norvig. Artificial Intelligence: A Modern Approach (4th ed). Harlow: Pearson, 2022.

Saballa, Joe. “US Army Seeking AI System that Predicts Enemy Actions.” The Defense Post (11 July 2023). https://www.thedefensepost.com/2023/07/11/us-army-ai-system/.

Saisubramanian, Sandhya, Shlomo Zilberstein, and Ece Kamar. “Avoiding Negative Side Effects Due to Incomplete Knowledge of AI Systems.” arXiv (18 October 2021). https://arxiv.org/pdf/2008.12146.pdf.

Scharre, Paul. Army of None. Autonomous Weapons and the Future of War. New York: W.W. Norton & Company, 2018.

Seshia, Sanjit A., Ankush Desai, Tommaso Dreossi, Daniel Fremont, Shromona Ghosh, Edward Kim, Sumukh Shivakumar, Marcell Vazquez-Chanlatte, and Xiangyu Yue. “Formal Specification for Deep Neural Networks,” University of California at Berkley, Electrical Engineering and Computer Sciences, Technical Report No. UCB/EECS-2018-25. 3 May 2018. https://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-25.pdf.

Shankar Siva Kumar, Ram, David O’Brien, Kendra Albert, Salome Viljoen, and Jeffrey Snover. “Failure Modes in Machine Learning Systems.” arXiv (25 November 2019). https://arxiv.org/ftp/arxiv/papers/1911/1911.11034.pdf.

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. “Membership Inference Attacks Against Machine Learning Models.” arXiv (31 March 2017): 1-16. https://arxiv.org/pdf/1610.05820.pdf.

Shumailov, Ilia, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. “Sponge Examples: Energy-Latency Attacks on Neural Networks.” arXiv (12 May 2021): 1-28. https://arxiv.org/pdf/2006.03463.pdf.

Stubbs, Austin. “LLM Hacking: Prompt Injection Techniques.” Medium (15 June 2023). https://medium.com/@austin-stubbs/llm-security-types-of-prompt-injection-d7ad8d7d75a3.

Toner, Helen and Ashwin Acharya. “Exploring Clusters of Research in Three Areas of AI Safety. Using the CSET Map of Science.” Center for Security and Emerging Technology. February 2022. https://cset.georgetown.edu/publication/exploring-clusters-of-research-in-three-areas-of-ai-safety/.

Truex, Stacey, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. “A Hybrid Approach to Privacy-Preserving Federated Learning.” AISec’19: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. November 2019. https://dl.acm.org/doi/10.1145/3338501.3357370.

United Nations Secretary-General. “Our Common Agenda. Policy Brief 9. A New Agenda for Peace.” July 2023. https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf.

United Nations General Assembly. “Report of the Group of Governmental Experts on Transparency and Confidence-Building Measures in Outer Space Activities.” A/68/189. 29 July 2013.

Ünver, H. Akın. “Digital Open Source Intelligence and International Security: A Primer.” Centre for Economics and Foreign Policy Studies. July 2018. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331638.

Urbina, Fabio, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. "Dual use of artificial-intelligence-powered drug discovery." Nature Machine Intelligence 4 (2022): 189-191. https://doi.org/10.1038/s42256-022-00465-9.

US Department of State. "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy." 16 February 2023. https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/.

US Senate - Armed Services. 118th Congress, 1st Session. "S.1394 - Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023." 1 May 2021. https://www.congress.gov/bill/118th-congress/senate-bill/1394/text.

US Department of Commerce, National Institute of Standards and Technology. "Artificial Intelligence Risk Management Framework." January 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Volodin, Sergei, Nevan Wichers, and Jeremy Nixon. "Resolving Spurious Correlations in Causal Models of Environments via Interventions." arXiv (9 December 2020): 1-13. https://arxiv.org/pdf/2002.05217.pdf.

Watts, Tom and Ingvild Bode. "Automation and Autonomy in Loitering Munitions Catalogue (v.1)." 25 April 2023. https://doi.org/10.5281/zenodo.7860762.

Wehsener, Alexa, Andrew W. Reddie, Leah Walker, and Philip J. Reiner. "AI-NC3 Integration in an Adversarial Context. Strategic Stability Risks and Confidence Building Measures." The Institute for Security and Technology. February 2023. https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

Wojton, Heather M., Daniel J. Porter, and John W. Dennis. "Test & Evaluation of AI-enabled and Autonomous Systems: A Literature Review." Institute for Defense Analyses. 9 March 2021. https://testscience.org/wp-content/uploads/formidable/20/Autonomy-Lit-Review.pdf.

UNIDIR

@unidir

/unidir

/un_disarmresearch

/unidirgeneva

/unidir

Palais de Nations
1211 Geneva, Switzerland

WWW.UNIDIR.ORG