AI, Autonomous Weapons and the Future of Strategic Stability and Arms Control

Artificial intelligence and autonomy are reshaping how militaries perceive threats, respond to crises, and manage the use of force. These technologies change the speed and scale at which information is processed and decisions are made. These advantages also introduce new uncertainties and vulnerabilities that have direct implications for strategic stability.

This factsheet examines how AI and autonomous weapons influence escalation dynamics, legal responsibility, arms racing, and the prospects for meaningful governance. It builds upon the foundations laid in the first factsheet and explains the policy consequences that follow from integrating AI into military operations.

AI, Escalation, and Crisis Instability

AI enabled systems promise faster decisions, better situational awareness, and the ability to interpret vast quantities of data. They support commanders by filtering sensor feeds, highlighting risks, and recommending possible courses of action. These tools seem to provide a clearer picture of fast moving situations, which is attractive in environments where reaction time is limited. However, the same systems that accelerate decisions also compress the time available for leaders to consider diplomatic alternatives or fully examine ambiguous signals. Faster analysis does not necessarily mean better judgement, and political leaders may be pushed toward quicker military choices before they fully understand unfolding events.

AI models used in military settings also suffer from technical limitations. They can misinterpret unfamiliar situations, become unreliable when conditions shift, and rely on correlations that are not well understood by human operators. Their outputs are often difficult to explain, even by developers, which creates opacity in high pressure environments. When these systems support threat assessment or early warning, opaque recommendations increase the risk of misreading an opponent’s behaviour. In crises where minutes matter, acting on a flawed assessment can trigger unnecessary escalation.

These technical shortcomings interact with human psychology. Operators may place too much confidence in machine recommendations, a tendency sometimes called automation bias. Over time, repeated reliance on automated analysis can erode human skill and make operators less able to challenge incorrect outputs. In a crisis, officials may defer to a system they believe to be more objective or precise. If the underlying data is biased or incomplete, this deference increases the risk of miscalculation. Because decision support flaws already distort interpretation of events, autonomy adds another layer in which the behaviour of systems themselves can become a source of escalation.

Autonomous military systems operate with varying degrees of independence once activated. They may sense, manoeuvre, or engage without direct human direction at every moment. When such systems behave unexpectedly in contested environments, it becomes unclear whether their actions reflect deliberate political choices, local tactical decisions, or technical failures. If an autonomous platform damages another state's forces, the target may not immediately know whether this was intended. This ambiguity undermines signalling in crises. Adversaries are more likely to assume worst case motives, increasing the chance that minor incidents escalate into wider confrontations.

Some states are exploring AI applications within nuclear command, control, and communications. While these uses are generally presented as decision support rather than automation of launch, they raise similar concerns about opacity, brittleness, and misinterpretation. The potential for rapid but poorly understood assessments to influence nuclear related decisions has led many experts to recommend caution, including avoiding highly autonomous systems in nuclear contexts and ensuring that humans remain the final authorities. Past discussions of automated nuclear response systems highlight the importance of avoiding situations where a software error or corrupted signal could produce catastrophic outcomes. These lessons underscore why carefully designed guardrails are necessary.

Together, these escalation pathways illustrate how AI and autonomy can alter crisis dynamics. They reduce warning time, introduce new failure modes, and blur the relationship between political intent and military action. These changes challenge assumptions that previously supported strategic stability and crisis management.

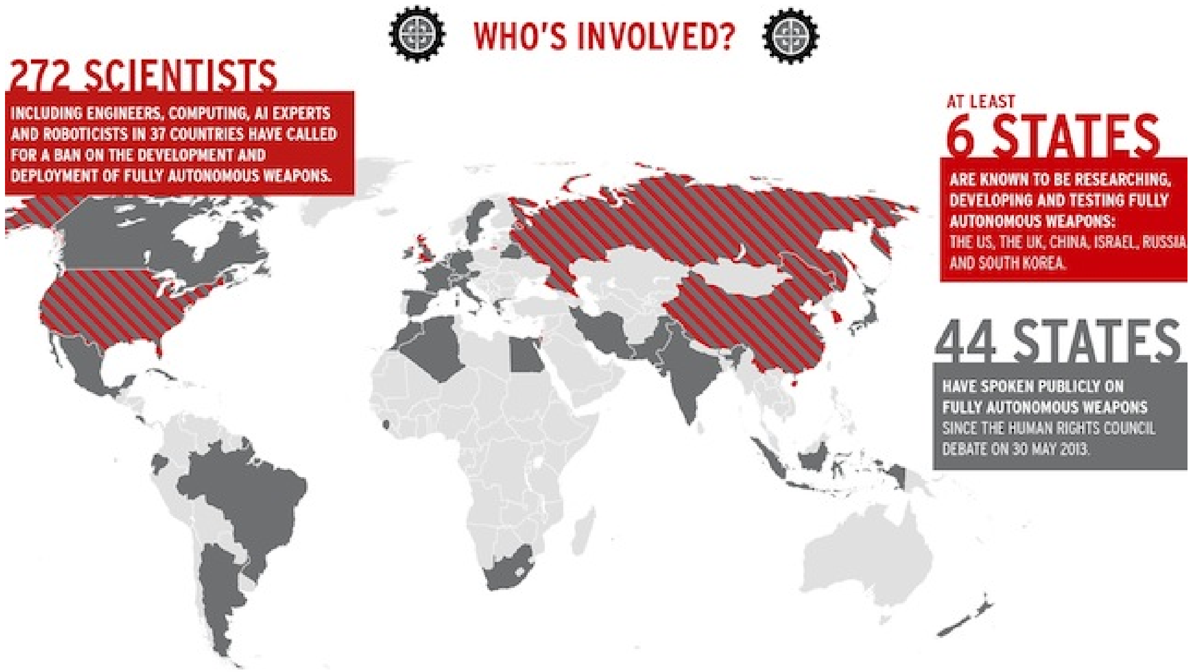
Legal, Ethical, and Accountability Challenges

International humanitarian law predates the widespread use of AI and autonomous weapons. Existing law regulates weapons that are physical and easily identified, yet provides little explicit guidance for decision making by algorithms or targeting systems that learn from data. In the absence of a specific treaty for AI based weapons, militaries rely on long standing principles such as distinction, proportionality, and military necessity. These principles require context sensitive judgement that AI systems currently struggle to perform.

Distinction requires separating combatants from civilians, while proportionality requires weighing anticipated military advantage against expected civilian harm. AI systems often rely on patterns in training data or on proxy indicators that stand in for complex human assessments. Because these challenges stem partly from the data and proxies AI systems use, concerns about bias and misclassification become central. AI can misidentify individuals who engage in civil defence functions or misinterpret routine civilian behaviour as threatening. Data drift can also degrade model performance over time, particularly in fluid conflict environments where behaviours evolve faster than datasets.

These typologies of error raise fundamental questions about accountability. If an autonomous system engages a target unlawfully, responsibility can become diffuse. Commanders, operators, engineers, and political leaders may all contribute to outcomes without clearly owning the decision. The opacity of advanced AI architecture complicates this further, since it becomes difficult to trace why a system reached a particular conclusion. Ethical arguments warn that transferring crucial aspects of lethal decision making to machines undermines long held expectations that humans must make moral assessments in war.

Existing legal frameworks still provide mechanisms to evaluate new technologies. States are required to review new weapons, means, or methods of warfare to ensure they comply with international law. These reviews are intended to identify foreseeable failures or harmful systemic behaviours. Since AI enabled systems can behave unpredictably, reviews must consider bias, training data quality, and the system's reliability in different scenarios. The Martens Clause reinforces that even when detailed rules are absent, the principles of humanity and public conscience remain applicable. Formal processes therefore already impose obligations to understand and mitigate risks before systems are deployed.



WHO'S INVOLVED?

272 SCIENTISTS

INCLUDING ENGINEERS, COMPUTING, AI EXPERTS AND ROBOTICISTS IN 37 COUNTRIES HAVE CALLED FOR A BAN ON THE DEVELOPMENT AND DEPLOYMENT OF FULLY AUTONOMOUS WEAPONS.

AT LEAST 6 STATES

ARE KNOWN TO BE RESEARCHING, DEVELOPING AND TESTING FULLY AUTONOMOUS WEAPONS: THE US, THE UK, CHINA, ISRAEL, RUSSIA AND SOUTH KOREA.

44 STATES

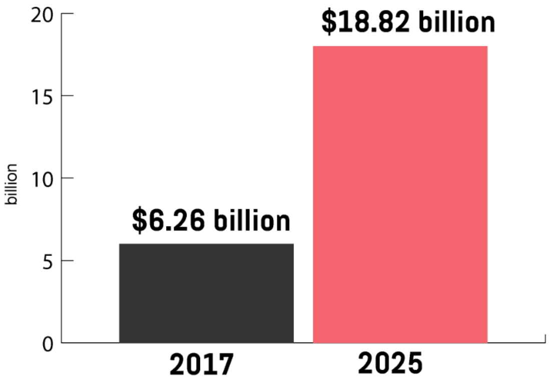
HAVE SPOKEN PUBLICLY ON FULLY AUTONOMOUS WEAPONS SINCE THE HUMAN RIGHTS COUNCIL DEBATE ON 30 MAY 2013.

Infographic from: https://www.stopkillerrobots.org/news/infographicaoav/

These obligations imply that meaningful human control is essential. Analysis of AI enabled targeting tools stresses that human oversight is necessary not only for technical safety but also as a legal requirement flowing from the duty to respect humanitarian law. Human involvement is needed to interpret context, understand intent, and ensure compliance with norms that algorithms cannot internalise. This conclusion holds particular weight in nuclear related decision processes, where the stakes make human judgement indispensable.

Implications for Arms Racing and Arms Control

AI and autonomous weapons influence military competition in powerful ways. Armed forces prize these technologies for their ability to improve sensing, coordinate complex operations, and support faster decisions. AI is also a dual use technology that is embedded across civilian and military sectors. This dual use nature complicates attempts to limit its spread, since the same techniques are essential for many peaceful applications.



Source: "Artificial Intelligence in Military Market by Offering (Software, Hardware, Services), Technology (Learning & Intelligence, Advanced Computing, AI Systems), Application (Information Processing, Cyber Security), Platform, Region - Global Forecast to 2025," *MarketsandMarkets*, 2018

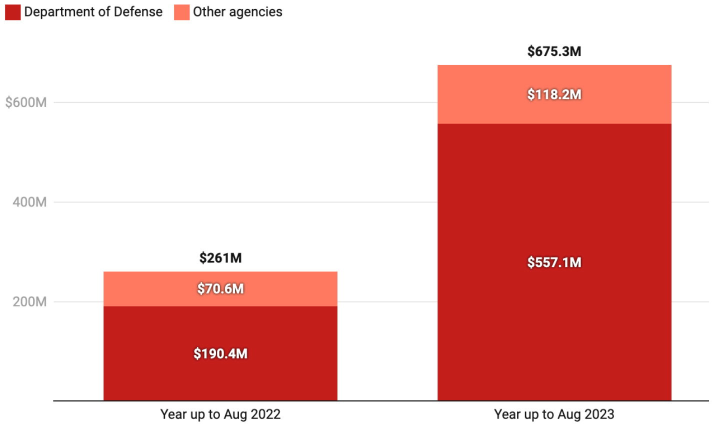
One major difficulty is definitional. AI is an enabling technology rather than a discrete weapon. Unlike chemical agents or ballistic missiles, it lacks a clear boundary that can be regulated. Different systems employ different models, datasets, and architectures, making it difficult for states to agree on a single category that should be limited. A broad slogan that AI should be banned does not resolve the practical challenge of identifying which functions are dangerous and which are beneficial.

Competitive incentives intensify the problem. States fear losing advantages in intelligence, targeting, and operational coordination. Unequal access to data, compute, and technical expertise already creates disparities. These disparities encourage rapid adoption and experimentation. Some analyses argue that, unless managed, this dynamic risks accelerating an unregulated race in military AI that leaves little space for cooperation or mutual restraint.

These pressures become more acute when AI enabled conventional systems intersect with nuclear forces. Enhanced surveillance and precision strike raise concerns that nuclear assets may become more vulnerable to disruption. This interaction can increase incentives to expand arsenals, diversify delivery systems, or adopt riskier postures during crises. The result is a feedback loop in which conventional AI capabilities influence nuclear stability, prompting reactions that contribute to arms racing.

U.S. military AI spending nearly tripled from 2022 to 2023

Total dollars obligated from AI-related federal contracts, $M



Contracts classified as AI-related if they had the term "artificial intelligence" or "AI" in the contract description.

Chart: Will Henshall for TIME • Source: Brookings Institute • Get the data • Created with Datawrapper

Verification adds another difficulty. Traditional arms control relies on inspecting physical systems or monitoring observable behaviour. AI systems are intangible, can be hidden in software, and can change rapidly through updates. Many potential verification methods are vulnerable to deception, and states are understandably reluctant to allow deep inspection of sensitive systems. Secrecy surrounding strategic programs further complicates transparency. Without effective verification, legally binding constraints become harder to negotiate and maintain.

Past arms control experience suggests that progress is most likely when agreements are narrowly tailored, easy to implement, and focused on specific risks rather than broad categories. Applied to AI, this approach points toward targeted limits on especially destabilising uses, such as automated nuclear launch procedures or certain applications of high autonomy weapons that are incompatible with legal or ethical standards. Narrower agreements avoid the pitfalls of attempting to control an entire technological field and instead address the clearest dangers.

Pathways to Governance and Risk Reduction

Although AI and autonomy present real challenges, there are practical steps that states can take to reduce risks. Confidence building measures, legal reviews, shared norms, and limited arms control provisions can all contribute to a safer environment.

Confidence building measures aim to improve communication, reduce uncertainty, and prevent accidents. One important area concerns testing and evaluation. Because system failures are a major source of risk, states can place greater emphasis on rigorous national testing requirements and share high level information about how they evaluate reliability and safety. This type of transparency does not require disclosing sensitive details, yet still promotes responsible practice. States may also coordinate on aspects of AI safety research, which supports more reliable systems across borders. These steps help clarify expectations and reduce the chance that unreliable models are used in sensitive roles.

A second group of measures focuses on autonomous systems operating in contested environments. To reduce the risk of unintended encounters, states could agree on behavioural rules or communication protocols for autonomous platforms. They could also consider marking systems to indicate their level of autonomy, providing clues about how they might behave in uncertain situations. Such measures help avoid misinterpretation and limit the likelihood that a malfunctioning platform is mistaken for a deliberate attack.

For nuclear systems, specialised confidence building measures are essential. These include commitments to preserve human decision making authority, avoid automated launch mechanisms, and refrain from deploying uninhabited systems designed to deliver nuclear weapons. Some proposals emphasise the need for independent reviews of nuclear decision processes to ensure that, even as AI is introduced for data analysis or threat assessment, final decisions remain firmly in human hands. These measures recognise that nuclear weapons present risks that cannot be managed through automation.

Legal and normative pathways provide further safeguards. Existing obligations under humanitarian law, including the requirement for weapon reviews, provide a basis for scrutinising AI enabled systems. These reviews assess whether a system can reliably comply with principles such as distinction and proportionality. They are also needed to identify and mitigate foreseeable model failures, including bias and data drift. Multilateral discussions on lethal autonomous weapons have begun to generate shared understandings about the need for human control, the importance of preventing discriminatory outcomes, and the application of existing law to new technologies. Political declarations and non binding codes of conduct can reinforce these commitments while more formal negotiations develop.

Finally, targeted arms control offers selective opportunities. Agreements that focus on particularly destabilising applications, such as automated nuclear response systems, are more feasible than broad attempts to regulate AI itself. States may also explore limits on deployment patterns or readiness postures of autonomous systems that pose crisis instability risks. By concentrating on specific hazards, rather than attempting to govern all military AI, arms control can meaningfully contribute to safety.

CONCLUSION

AI and autonomy are altering the foundations of strategic stability. They increase the speed of decision making, introduce new technical failure modes, and create ambiguity about intent. At the same time, dual use characteristics and competitive pressures make broad control difficult. Yet the same qualities that make AI attractive for military use also make governance essential. By strengthening testing and evaluation, improving transparency, reinforcing human control, and pursuing targeted limits on the most dangerous applications, states can reduce the risks that accompany AI enabled warfare while preserving pathways for responsible use.