UNIDIR

Agentic AI and International Security: Navigating the Shift Towards Agency

GIACOMO PERSI PAOLI



Acknowledgements

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. This report was prepared by the Centre of Excellence on AI, Peace and Security, which is funded by the Governments of France, Germany, Italy, the Kingdom of the Netherlands, Pakistan, the Republic of Korea and Switzerland, and by Microsoft and the Fondation pour l'adaptation de la Genève internationale. Additional support for this report was received by the Friedrich-Ebert-Stiftung.

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.

About the Centre of Excellence on AI, Peace and Security

The Centre of Excellence on AI, Peace and Security is UNIDIR's global, permanent platform dedicated to advancing AI governance in the peace and security domain. Based in Geneva, it brings together research, dialogue and capacity-building to support more effective and inclusive approaches to the challenges AI presents.

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 authors. They do not necessary reflect the views or opinions of the United Nations, UNIDIR, its staff members or sponsors.

Citation

Giacomo Persi Paoli. *Agentic AI and International Security: Navigating the Shift Towards Agency*. Geneva: UNIDIR, 2026. https://doi.org/10.37559/SECTEC/26/AI/09.

Cover image: AI agentic workflow automation diagram featuring interface nodes, triggers, and data tool dashboards rendered in 3D. By Garapic

About the author



Giacomo Persi Paoli

Giacomo Persi Paoli is a UNIDIR Fellow and served as Head of UNIDIR's Security and Technology Programme from 2019 to 2026. Dr Persi Paoli specializes in the implications of emerging technologies – particularly artificial intelligence and cyber – for international peace and security. His recent work focuses on the governance of these transformative technologies, including their impact on arms control, strategic stability and global security frameworks. Dr Persi Paoli holds a doctorate in economics from the University of Rome, a master's degree in political science from the University of Pisa and a master's degree in international affairs and diplomacy from the University of Trieste.

Acronyms and abbreviations

AI	Artificial intelligence
C2	Command and control
LAWS	Lethal autonomous weapon systems
LLM	Large language model
OODA	Observe–orient–decide–act
REAIM	Responsible AI in the Military Domain
TEVV	Testing, evaluation, verification and validation
UAV	Unmanned aerial vehicle

Table of contents

Executive summary **6**

1. Introduction **9**

1.1 Scope and limitations 10

1.2 Structure of this report 10

2. What is agentic AI? **11**

2.1 Key features of agentic AI 12

2.2 Agentic AI versus rule-based systems 12

3. Goal and constraint modelling **14**

3.1 How are goals modelled? 14

3.2 How are constraints modelled? 15

3.3 Misalignment and control failure 17

4. Agentic AI for cybersecurity **19**

5. Agentic AI in military command and control **23**

6. Agentic AI in autonomous physical systems **26**

7. Preliminary mapping of risks **29**

7.1 Cross-cutting risks of agentic AI 29

7.2 Domain-specific risk snapshots 31

7.3 Systemic and strategic risks 33

8. Foundations for the governance of agentic AI **36**

8.1 Building on existing governance frameworks 37

8.2 Agent-to-agent interaction: An emerging international governance challenge 38

8.3 Priorities for international cooperation 39

9. Conclusion **41**

Executive summary

Artificial intelligence (AI) is already becoming increasingly embedded across military and security activities, but the emergence of **agentic AI** introduces a potentially important change in the relationship between humans and AI systems. Agentic systems can pursue objectives over multiple steps, determine intermediate actions, use tools, adapt their behaviour and take actions with reduced human intervention. Their significance for international peace and security therefore lies not simply in greater autonomy, but in a **changing distribution of operational discretion between humans and AI systems**: humans may increasingly specify objectives and boundaries while AI systems determine more of the intermediate actions through which those objectives are pursued. The term "agency" is used here in a functional sense and should not be confused with human agency, consciousness, intention or legal responsibility.

The maturity of these capabilities remains uneven. The clearest current evidence of agentic capability is found in software and cyber environments, where systems can already undertake increasingly complex sequences of actions using external tools. Applications in military command and control (C2) and autonomous physical systems are less mature and remain largely experimental. This distinction is important: **advanced autonomy should not automatically be equated with agency**, and experimental capability should not be presented as evidence of widespread operational deployment. This report therefore distinguishes throughout between demonstrated agentic capabilities, experimental systems, enabling technologies and prospective applications.

How goals, permissions and constraints are specified becomes particularly important as more of the pathway from objective to action is delegated to a system. Agentic systems cannot be made safe simply by providing them with a high-level instruction and assuming that they will interpret it as intended. Objectives may be incomplete or ambiguous; a measurable proxy for a goal may diverge from the underlying goal; and capable systems can sometimes discover unintended ways of satisfying an evaluation or reward mechanism. Effective control consequently depends not only on behavioural instructions but also on **technically enforceable permissions and constraints**, including restrictions on tools and data, hard-coded rules, human approval gates and independent monitoring. In high-stakes security applications, an additional difficulty arises because legal and policy requirements cannot always be translated into fixed computational rules. Context-dependent legal judgments, including those required by international humanitarian law, remain the responsibility of humans and States.

Three application areas illustrate both the potential significance of agentic AI and the uneven state of development: cybersecurity, military C2 and autonomous physical systems.

Cybersecurity currently provides the clearest evidence base. Agentic systems can connect analysis to action by querying telemetry, inspecting systems, invoking security tools, testing hypotheses and adapting subsequent steps. Emerging applications include threat hunting, incident investigation, active defence, cyber deception, vulnerability discovery and recovery. These capabilities are inherently dual-use: the same ability to navigate systems, discover vulnerabilities and adapt over multiple steps can support defensive or offensive activity. Recent evaluations also demonstrate that increasingly capable agents can perform longer cyberattack chains with limited human intervention, although advanced end-to-end operations still generally require human involvement.

In **military command and control**, the potential shift is from AI that primarily supports analysis and recommendation towards systems that may be authorized to pursue objectives across multiple stages of a C2 process and implement bounded actions. Possible applications include planning support, adaptive coordination, resilience under degraded communications and simulation or war-gaming. Experimental systems already demonstrate elements of multi-agent planning, command execution and replanning in simulated environments, but there remains limited public evidence of fully agentic systems performing operational C2 functions in real-world military settings. The key governance question is therefore not whether AI can contribute to C2 – it already does – but which functions and decisions may appropriately be delegated, under what conditions, and with what mechanisms for supervision, intervention and accountability.

For **autonomous physical systems**, agency could extend existing autonomy from particular functions towards mission-level planning and adaptation. Autonomous platforms already perform tasks such as navigation, sensing, coordination and targeting without continuous human intervention. Agentic AI could connect many of these functions by allowing a system to interpret higher-level objectives, determine intermediate steps, revise plans and coordinate with other systems over the course of a mission. Current evidence nevertheless suggests that caution is required: many operational examples demonstrate advanced autonomy rather than agency, while stronger evidence of agentic behaviour comes primarily from experimental or simulated systems. The distinction is particularly important in relation to lethal autonomous weapon systems (LAWS). Autonomous targeting and agentic AI are overlapping but different concepts; a weapon may autonomously identify, select and engage targets without possessing broader capabilities for goal-directed replanning or mission adaptation.

Agentic AI does not create an entirely new set of risks, but it can **amplify existing AI risks and alter how they translate into real-world effects**. An erroneous output from a conventional system may be presented to a human for consideration; an erroneous output in an agentic workflow may instead become an input to further planning, tool use and action. Cross-cutting concerns include misalignment of goals and specifications, failures of reliability and propagation of errors, circumvention of constraints, adversarial manipulation, multi-agent interaction, difficulties in traceability and accountability, and escalatory dynamics. Again, the evidence base is uneven. Specification gaming, prompt injection, manipulation of oversight mechanisms and other forms of agent failure have already been observed in controlled evaluations, while more severe scenarios involving autonomous military escalation or large-scale loss of human control remain theoretical.

At the systemic level, three issues merit particular attention. First, more capable and persistent systems may make **loss of effective human control** harder to detect or reverse. However, current systems do not possess the combination of capabilities associated with severe loss-of-control scenarios. Second, strategic and commercial competition may create **speed-versus-assurance pressures**. These can encourage deployment before testing, institutional adaptation and oversight mechanisms have matured. Third, because many agentic capabilities depend on general-purpose models, commercial software and widely available tools, they may **diffuse more rapidly** than traditional military technologies and become accessible to a wider range of State and non-State actors.

These developments do not occur in a governance vacuum. International law continues to apply, and existing initiatives on responsible military AI already address principles such as human responsibility,

defined uses, testing, assurance and safeguards. The immediate governance challenge is therefore **not necessarily to create a separate international regime for agentic AI**, but to determine whether existing frameworks remain adequate as systems gain greater capacity to plan, act, adapt and interact. This requires the stress-testing of existing commitments against future action, tool use, delegated authority and interaction with other agents, while translating high-level principles into practical controls to govern permissions, monitoring and conditions for human intervention.

One issue may warrant particularly early international attention: **persistent interactions between agents operated by different actors**. Many international norms, confidence-building measures and crisis-management mechanisms were developed on the assumption that consequential signalling, interpretation and restraint are mediated by humans. Those rules and obligations continue to apply to States when AI is involved, but it cannot simply be assumed that autonomous agents will reproduce the behaviours on which these arrangements depend. States may therefore need to develop shared expectations concerning agent identity and authentication, delegated authority, responses to ambiguous behaviour, conditions for human referral or termination, and mechanisms for recording and attributing actions. Political agreement should precede technical implementation: international norms and rules would continue to govern States, while technical protocols could help ensure that agents acting on their behalf operate consistently with those expectations.

Several priorities for international cooperation can be identified: developing a sufficiently shared understanding of agency; incorporating agentic behaviour into assurance and risk-management processes; using scenario exercises and technical-policy dialogue to examine agent-to-agent interaction; strengthening national implementation capacity; and maintaining inclusive multi-stakeholder engagement, particularly because much of the relevant technology and technical expertise sits outside government.

The central policy task is therefore **anticipatory rather than reactive**. Agentic AI does not remove humans or States from responsibility for decisions affecting international peace and security. Instead, it makes the conditions under which authority is delegated, bounded, supervised and reclaimed increasingly important. Early engagement can help identify where existing governance remains sufficient, where implementation needs to evolve, and where emerging forms of interaction may require new political understandings before operational practices become entrenched.



Artificial intelligence and deep learning technology. By The 2R Artificiality

1. Introduction

Artificial intelligence (AI) is becoming increasingly embedded across the military and security domain, including in intelligence analysis, cyber operations, decision support, logistics and autonomous systems.[1] These applications vary significantly in their capabilities and in the extent to which they support, shape or substitute for human decision-making and action.[2]

The emergence of **agentic AI** introduces a potentially important new dimension. Rather than simply generating information or recommendations, agentic systems can pursue objectives over multiple steps and take actions with reduced human intervention. For international peace and security, their significance lies in a **changing distribution of operational discretion between humans and AI systems**: not only whether a system can perform a task autonomously, but how much of the pathway from an objective to action it can determine for itself.[3]

This creates opportunities as well as risks. Agentic systems can support faster and more persistent operations, can coordinate activity across multiple systems, and can operate in environments where continuous human attention is difficult. At the same time, greater discretion, poorly specified objectives or misunderstood constraints in an AI system can amplify mistakes by allowing them to propagate across sequences of actions. Agency may also complicate existing approaches to human oversight, responsibility and control.

---

1 General Assembly, resolution 80/58, "Artificial intelligence in the military domain and its implications for international peace and security", 1 December 2025, https://docs.un.org/A/RES/80/58.

2 UNIDIR Security and Technology Programme, *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security: An Evidence-Based Road Map for Future Policy Action* (Geneva: UNIDIR, 2025), https://unidir.org/publication/artificial-intelligence-in-the-military-domain-and-its-implications-for-international-peace-and-security-an-evidence-based-road-map-for-future-policy-action/.

3 Organisation for Economic Co-operation and Development (OECD), *The Agentic AI Landscape and Its Conceptual Foundations*, OECD Artificial Intelligence Papers no. 56 (Paris: OECD Publishing, 2026), https://doi.org/10.1787/396cf758-en.

The maturity of these capabilities remains uneven. The evidence reviewed in this report suggests that the clearest current examples of agentic capability are found in software and cyber environments, while applications in military command and control (C2) and autonomous physical systems remain more experimental. Advanced autonomy should therefore not automatically be equated with agentic AI, nor should experimental capability be treated as evidence of widespread operational deployment.

Nevertheless, the issue warrants early policy attention. Agentic capabilities are developing rapidly within predominantly civilian and dual-use technology ecosystems and may therefore diffuse more quickly than traditional military technologies. The central question is **how agentic AI may alter existing challenges** concerning human control, reliability, accountability, escalation and compliance with international law. Of particular importance is the prospect of increasingly autonomous interactions between agents operated by different actors, an issue for which existing international norms, confidence-building measures and crisis-management arrangements were not originally designed.

1.1 Scope and limitations

Agentic AI remains a rapidly evolving field, with terminology and technical capabilities still developing. This report does not seek to provide a definitive taxonomy or assume that the capabilities discussed are already widely deployed in military and security settings. Instead, it uses agency as a functional lens through which to examine emerging systems capable of pursuing goals, planning, acting, adapting and interacting with reduced human intervention.

The report is intended as an introductory primer for policymakers. Its purpose is to establish a common conceptual foundation, to distinguish demonstrated capabilities from experimental and prospective applications, and to identify issues that may warrant further international attention as the technology evolves.

1.2 Structure of this report

The report moves from concepts to applications, risks and governance. **Section 2** introduce agentic AI and **Section 3** examines how goals, permissions and constraints shape its behaviour. **Section 4** explores applications in cybersecurity, **Section 5** those in military command and control, and **Section 6** autonomous physical systems. **Section 7** provides a preliminary mapping of cross-cutting and systemic risks. **Section 8** examines how existing governance frameworks can be applied to agentic AI and identifies priorities for international cooperation, including the emerging challenge of agent-to-agent interaction. The report concludes in **Section 9**.

2. What is agentic AI?

Agentic AI refers to AI systems that can pursue specified objectives over multiple steps, determine how to achieve them and take actions with limited human intervention.[4] These systems typically combine capabilities such as reasoning and planning with memory or maintained context along with access to tools that allow them to interact with digital or physical environments.[5]

Agency should be distinguished from autonomy. Autonomy describes the degree to which a system can operate without direct human involvement; agency concerns its capacity to pursue goals, determine intermediate actions and adapt its behaviour in response to changing circumstances.[6] A system can therefore operate autonomously without necessarily exhibiting the broader characteristics associated with contemporary agentic AI. Agency in AI should also be distinguished from human agency (see Box 1).

BOX 1

"Agency" in AI versus human agency

In this report, the term **agency** is used in a functional or engineering sense. It refers to the capacity of an AI system to pursue an assigned objective, select and sequence actions, interact with its environment, and adapt its behaviour with limited human intervention.

This should not be confused with **human agency**. Describing an AI system as "agentic" does not imply that it possesses consciousness, subjective experience, intrinsic intentions, moral judgement or free will.[7] Nor does the ability to plan, use tools or act autonomously by itself establish agency in the stronger sense relevant to moral or legal responsibility.

Current agentic systems ultimately operate with objectives and capabilities and in environments established by humans. Their significance lies in **how much operational discretion humans grant to AI systems**, including the ability to determine and take intermediate actions in pursuit of an assigned objective.

For policymakers, the crucial development is therefore not simply that these systems are more autonomous, but that they can plan and act on their environment and – increasingly – that they can interact with

4 OECD, *The Agentic AI Landscape and Its Conceptual Foundations*.

5 Mohamad Abou Ali, Fadi Dornaika and Jinan Charafeddine, "Agentic AI: A comprehensive survey of architectures, applications, and future directions", *Artificial Intelligence Review*, vol. 59 (2026), article 11, https://doi.org/10.1007/s10462-025-11422-4.

6 OECD, *The Agentic AI Landscape and Its Conceptual Foundations*.

7 Peter Kahl, "How continuity distinguishes autonomy from agency in agentic AI", *Discover Artificial Intelligence*, vol. 6 (2026), article 1187, https://doi.org/10.1007/s44163-026-01675-5.

other agents. The growing availability of tools that allow AI agents to modify external environments illustrates this shift from systems that primarily generate information towards systems capable of taking action.[8]

2.1 Key features of agentic AI

There is no single universally accepted set of characteristics that defines agentic AI, and systems can display different degrees of agency.[9] Nevertheless, several features occur across recent definitions and technical literature.

Agentic systems are **goal-directed**: they operate in pursuit of an objective, rather than simply producing a response to an isolated input. More capable agents can break down complex objectives into intermediate tasks and determine the sequences of actions that are needed to achieve them.[10]

They exhibit **planning and action selection**: they can choose what to do next on the basis of available information and the results of previous actions, rather than following a fully predetermined sequence.

They can also display **adaptability**: they can revise plans or select alternative actions when circumstances change or an initial approach fails.[11] This should not necessarily be understood as "learning" since many agents adapt during execution without modifying the underlying AI model.

Agentic systems may also display **persistence**: they pursue objectives across multiple steps or over extended periods with limited human intervention. Many contemporary architectures additionally use memory or maintained context to preserve relevant information across actions and interactions, although persistent memory is not a necessary feature of every agent.[12]

For a military audience, these capabilities can be understood as spanning functions analogous to the **observe–orient–decide–act (OODA) loop**: agents can receive information from an environment, interpret it, select or plan actions, execute them through available tools or systems, and use the resulting feedback to inform subsequent action. This is a useful functional analogy, rather than a description of a specific technical architecture.

2.2 Agentic AI versus rule-based systems

A useful way of illustrating agency is to contrast contemporary agentic systems with traditional **rule-based automation**, while recognizing that the distinction is not absolute (see Table 1). Rule-based systems follow execution pathways specified largely in advance – for example, rules in the form "if X occurs, then perform Y". They can operate autonomously, but their range of possible responses is principally

---

8 Merlin Stein, "How are AI agents used? Evidence from 177,000 MCP tools", *British AI Security Institute*, 2026, https://www.aisi.gov.uk/research/how-are-ai-agents-used-evidence-from-177-000-mcp-tools.

9 OECD, *The Agentic AI Landscape and Its Conceptual Foundations*.

10 Abou Ali et al., "Agentic AI".

11 Abou Ali et al., "Agentic AI".

12 Abou Ali et al., "Agentic AI".

determined by predefined rules. Agentic systems are instead given an objective, available actions and constraints, and can determine more of the sequence of steps used to pursue that objective.

The distinction therefore lies less in whether a system is automated or autonomous than in how much of the pathway from objective to action is determined in advance. In rule-based automation, developers specify much of the permissible execution pathway. In agentic systems, developers and operators increasingly specify goals, permissions, constraints and conditions for human intervention, while leaving the system greater flexibility to determine intermediate actions.[13] Agentic architectures can nevertheless incorporate rule-based components or hard technical constraints: agency does not imply the absence of rules.

TABLE 1

Rule-based automation versus agentic AI

Feature	Rule-based automation	Agentic AI
Primary control logic	Predefined rules and workflows	Goal-directed planning within defined permissions and constraints
Execution pathway	Largely specified in advance	Determined dynamically over multiple steps
Action selection	Executes predefined responses	Selects and sequences available actions
Adaptation	Limited to anticipated conditions and branches	Can revise plans or actions in response to feedback and changing circumstances
Persistence	Executes a bounded workflow or response	May pursue objectives across multiple steps or extended periods
Human instruction	Humans largely determine the pathway in advance	Humans specify objectives and boundaries; the system determines more of the intermediate pathway
Interaction	Typically operates through predefined interfaces	Can interact dynamically with tools, environments, humans and other agents

13 Infocomm Media Development Authority of Singapore, "Model AI governance framework for agentic AI", updated May 2026, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/updated-model-ai-governance-framework-for-agentic-ai.

3. Goal and constraint modelling

The effectiveness and safety of agentic AI systems depend heavily on how their goals, permissions and constraints are specified. Because these systems can determine intermediate steps and take actions with limited human intervention, developers cannot prescribe every possible sequence of behaviour. Instead, they define objectives, available actions, permissions and boundaries within which the system is expected to operate.[14] Ensuring that the resulting behaviour remains consistent with human intent is one of the central challenges in developing and deploying agentic systems safely.

Building on the distinction in Section 2, a useful overview is to think of the move from rule-based automation to agentic systems as a shift from **"modelling in" acceptable behaviour** (i.e., specifying acceptable behaviour in the model) to increasingly **"modelling out" unacceptable behaviour**. Because agentic systems have greater flexibility to determine how an objective is achieved, developers must place greater emphasis on defining the boundaries within which they may act. This is only an approximation: safe agentic systems can still incorporate predefined workflows, hard-coded rules, restricted action spaces and deterministic safeguards.[15]

The shift nevertheless has important implications:

- Developers must consider failure modes and unintended strategies that cannot all be enumerated in advance.
- Evaluation must test behaviour across diverse and changing scenarios, rather than only predefined inputs and outputs.
- Oversight increasingly requires assessing not only whether an action was anticipated, but whether it remained within the system's assigned objective, authority and constraints.[16]

3.1 How are goals modelled?

Goals can enter an agentic system at different stages and in different forms. It is useful to distinguish objectives used to train or optimize a system from the goals that an agent is instructed to pursue when deployed.

- **Reward or optimization objectives.** In reinforcement-learning and optimization-based systems, behaviour can be shaped by reward or objective functions that assign value to particular outcomes. These functions may balance multiple considerations, such as task performance, resource use or penalties for undesirable actions. Importantly, these objectives often shape how a system is trained or optimized, rather than serving as the explicit instruction given to a deployed agent.
- **Natural-language goals.** Contemporary agents based on large language models (LLMs) commonly receive objectives through natural-language instructions, such as "identify and contain malicious

14 OECD, *The Agentic AI Landscape and Its Conceptual Foundations*.

15 Infocomm Media Development Authority of Singapore, "Model AI governance framework for agentic AI".

16 US National Institute of Standards and Technology (NIST), "Practices for automated benchmark evaluations of language models", NIST AI 800-2, Initial Public Draft, January 2026, https://doi.org/10.6028/NIST.AI.800-2.ipd.

activity without disrupting legitimate services". Natural language makes it possible to specify complex objectives flexibly, but it can also introduce ambiguity concerning priorities, exceptions and how conflicting considerations should be resolved.[17]

- **Structured or symbolic goals.** Agentic systems can also represent objectives as desired states, logical conditions or structured planning problems. Symbolic and hybrid architectures may reason from a current state towards a defined goal state using available actions and explicit preconditions, while contemporary systems increasingly combine such approaches with generative AI.[18]

In practice, these mechanisms can coexist. A deployed agent may receive a natural-language mission, decompose it into intermediate goals, invoke an optimization routine for a particular task and operate within permissions established by external software.

The way a goal is specified matters because optimization can expose gaps between a measurable objective and the designer's actual intent. **Specification gaming** occurs when a system satisfies the literal specification or proxy used to measure success while failing to achieve the intended outcome.[19] More capable agents may have more opportunities to discover such unintended strategies because they can plan over multiple steps, use tools and modify their environment.[20]

3.2 How are constraints modelled?

If goals describe what an agent should achieve, constraints define the conditions under which it may pursue those goals. For agentic systems, constraints are particularly important because natural-language instructions alone do not necessarily constitute enforceable technical boundaries (see Box 2).

Constraints can operate at several levels:

- **Action and permission constraints.** These restrict the tools, data, networks or physical capabilities available to the agent. An agent might, for example, be permitted to inspect files but not delete them, or to recommend a configuration change but not implement it.
- **Hard-coded rules and policy filters.** These prevent particular actions or require specified conditions before they may occur.
- **Optimization constraints and penalties.** These discourage undesirable outcomes by incorporating them into reward or objective functions.
- **Human approval gates.** These require authorization before consequential or irreversible actions are executed.
- **External monitoring and safety layers.** These can inspect proposed actions, block prohibited behaviour or terminate the system independently of the agent itself.

---

17 OECD, *The Agentic AI Landscape and Its Conceptual Foundations*.

18 Abou Ali et al., "Agentic AI".

19 See, for example, Maia Hamin and Benjamin Edelman, "Cheating on AI agent evaluations", NIST Center for AI Standards and Innovation, 28 November 2025, updated 2 December 2025, https://www.nist.gov/caissi/cheating-ai-agent-evaluations.

20 Arun Morampudi et al., "A survey of reward hacking in agentic large language model systems", *Discover Artificial Intelligence*, vol. 6 (2026), article 825, https://doi.org/10.1007/s44163-026-01980-z.

BOX 2

Translating legal and normative constraints

Security applications of agentic AI create an additional challenge because some relevant constraints derive from legal and policy frameworks that cannot easily be reduced to fixed computational rules (see Table 2).

International humanitarian law provides a clear example. Rules governing distinction, proportionality and precautions in attack require context-specific judgments by those who plan, decide upon and carry out attacks. Proportionality, for example, requires assessing whether expected civilian harm would be excessive in relation to the concrete and direct military advantage anticipated; it is not simply the application of a universal numerical threshold. Similarly, recent analysis cautions against assuming that the qualitative and contextual judgments involved in proportionality can be reduced to quantifiable inputs for automated decision-making.[21]

This does not mean that legal requirements cannot inform technical constraints. Systems can be restricted by factors such as geographical area, duration, target categories, available actions or requirements for human authorization. Such limits may help reduce the circumstances in which unlawful outcomes can arise. They should not, however, be confused with encoding the legal rule itself.

The distinction is important for agentic AI. A technical safeguard can prevent an agent from performing a specified action, but abstract requirements (e.g., proportionality, necessity or appropriate treatment of persons whose status changes during an operation) may depend on contextual judgments that cannot be fully captured through predetermined rules. Human responsibility for applying international law therefore remains essential even as technical controls become more sophisticated.[22]

These approaches differ in strength. The effectiveness of a natural-language instruction such as "do not access external networks" depends partly on the agent correctly interpreting and following that instruction. Removing network access at the system level creates a much stronger technical constraint. Current governance guidance for agentic systems therefore increasingly emphasizes bounding agents' access to tools and data, white-listing permitted services and requiring human approval for higher-risk actions.[23]

Testing and red-teaming complement these controls but should not be treated as constraints themselves. Their purpose is to discover whether instructions, permissions or safeguards fail under adversarial or unexpected conditions. Controlled evaluations of tool-using AI agents have documented cases where a system has exploited gaps between the intended objective and the evaluation mechanism (e.g., skipping

21 Jessica Dorsey, "The erosion of human(e) judgement in targeting? Quantification logics, AI-enabled decision support systems and proportionality assessments in IHL", *International Review of the Red Cross*, no. 930 (2025), https://doi.org/10.1017/S1816383125100969.

22 International Committee of the Red Cross, "FAQ: Artificial intelligence in the military domain", 11 June 2026, https://www.icrc.org/en/article/faq-artificial-intelligence-in-military-domain.

23 Infocomm Media Development Authority of Singapore, "Model AI governance framework for agentic AI".

required verification steps, exploiting task-adjacent information or tampering with evaluation-relevant functions). These illustrate why compliance with an instruction cannot simply be assumed.[24] Evaluation of agentic systems must therefore examine both outcomes and the pathways through which those outcomes are achieved.[25]

3.3 Misalignment and control failure

Even where goals and constraints are carefully specified, agentic systems can produce outcomes that diverge from what their users intended. The term **misalignment** is often used broadly, but several distinct mechanisms should be differentiated.

- **Goal or specification misalignment.** The objective provided to the system may imperfectly represent the user's actual intent. For example, an instruction to "maximize system uptime" could create incentives to postpone disruptive but necessary security updates if longer-term resilience is not captured in the objective.
- **Reward hacking and specification gaming.** A system may discover strategies that improve the metric, reward or evaluator used to measure success without accomplishing the underlying task. Recent research documents agentic systems exploiting test environments, modifying evaluation mechanisms and manipulating features used to determine success.[26] Agentic deployment can make this problem more consequential because, instead of merely producing an undesirable output, systems may be able to alter files, logs, tools or other elements of the environment.[27]
- **Conflicting objectives and constraints.** Agents may be asked simultaneously to pursue objectives that cannot all be maximized simultaneously (e.g., speed, coverage, resource conservation and risk minimization). Unless priorities and acceptable trade-offs are sufficiently clear, the system may resolve these tensions in ways its operators did not anticipate.

These forms of misalignment should be distinguished from broader reliability and control failures. A system may fail because it encounters unfamiliar conditions, while a capable agent may also circumvent monitoring or other mechanisms intended to constrain its behaviour. In agentic systems, such failures can become more consequential because the result of one action can inform subsequent planning and behaviour. (Section 7 examines these broader risk pathways and their implications for international security.)

For goal and constraint modelling, the practical implication is that objectives alone are insufficient. They must be accompanied by enforceable permissions and constraints, together with testing capable of evaluating behaviour across extended sequences of action.

---

24 Kunvar Thaman, "Reward hacking benchmark: Measuring exploits in LLM agents with tool use", *Proceedings of the 43rd International Conference on Machine Learning*, 22 September 2026, https://openreview.net/forum?id=YMA0ByEdVj.

25 US National Institute of Standards and Technology (NIST), "Practices for automated benchmark evaluations of language models".

26 Thaman, "Reward hacking benchmark".

27 Morampudi et al., "A survey of reward hacking in agentic large language model systems".

TABLE 2

Challenges in goal and constraint modelling for security contexts

Challenge	Description
Goal ambiguity	Broad objectives may leave important priorities, exceptions or trade-offs unspecified
Proxy problems	A measurable proxy for success may diverge from the underlying objective
Context dependence	Legal, ethical or operational constraints may depend on information that cannot be fully specified in advance
Constraint enforceability	Behavioural instructions can be interpreted or circumvented unless supported by technical limits on permissions and actions
Environmental shift	Goals and constraints established under one set of assumptions may produce inappropriate behaviour when conditions change
Long-horizon verification	As agents take longer sequences of adaptive actions, testing whether behaviour remains aligned with objectives and constraints becomes increasingly difficult

The first three sections of this report provide an overarching introduction to the technology. The following three examine specific use cases relevant to international peace and security: cybersecurity, command and control, and autonomous physical systems.

4. Agentic AI for cybersecurity

Cybersecurity is one of the most natural domains for the application of agentic AI. Cyber operations take place in digital environments that are data-rich, highly instrumented and accessible through software tools and interfaces. Many cybersecurity tasks also involve iterative cycles of observation, analysis, decision and action that map closely onto the way agentic systems operate. Agentic AI could therefore help defenders process large volumes of security data, investigate threats, coordinate across different security tools, and respond more rapidly and continuously to incidents.[28]

Agentic AI extends existing AI-enabled cybersecurity tools by coupling analytical capabilities with the ability to plan and take action.[29] Depending on their design and permissions, agents can query network telemetry, inspect files and systems, invoke cybersecurity tools, generate and test hypotheses, execute actions, and revise their approach based on the results. These characteristics make them relevant to functions such as threat hunting, alert triage, incident response, vulnerability management and recovery.[30] Early operational deployments and experimental systems are already demonstrating elements of these capabilities, including agents that autonomously triage alerts, investigate suspicious activity, analyse malware and coordinate multi-stage security investigations.[31]

These characteristics create opportunities for cyber defence but can inherently be adapted to cyber offense. The same capabilities that allow an agent to search for vulnerabilities, navigate networks, use cyber tools and adapt its approach can also support reconnaissance, exploitation and other offensive cyber activity. Recent evaluations suggest that frontier AI agents are becoming increasingly capable of completing longer sequences of cyberattack tasks with limited human intervention. However, substantial limitations remain and advanced end-to-end cyber operations still generally require human involvement.[32]

The applications of agentic AI to cybersecurity can be grouped into several core functions that illustrate its potential value compared with more traditional automated tools.

- **Autonomous threat hunting and investigation.** Agentic systems can continuously analyse security data, investigate anomalies and pursue hypotheses across multiple systems and data sources.

---

28 Nir Kshetri, "Transforming cybersecurity with agentic AI to combat emerging cyber threats", *Telecommunications Policy*, vol. 49, no. 6 (2025), 102976, https://doi.org/10.1016/j.telpol.2025.102976; Hamed AlQahtani and Paras Ahuja, "Secure autonomous cyber defense with LLM agents: A systematic review of autonomy, tool-augmented reasoning, and governance constraints", *Computers & Electrical Engineering*, vol. 135 (2026), 111184, https://doi.org/10.1016/j.compeleceng.2026.111184.

29 For an overview of the AI-cybersecurity nexus, see: Giacomo Persi Paoli and Samuele Dominioni. *Exploring the AI-ICT Security Nexus*. Geneva: UNIDIR, 2024. https://unidir.org/publication/exploring-the-ai-ict-security-nexus/

30 AlQahtani and Ahuja, "Secure autonomous cyber defense with LLM agents".

31 Google Cloud, "Agentic AI use case: Orchestrate security operations workflows", Cloud Architecture Center, last reviewed 8 April 2026, https://docs.cloud.google.com/architecture/agentic-ai-orchestrate-security-ops-workflows; Microsoft, "Deploy AI agents in Microsoft Defender", Microsoft Learn, updated 5 August 2026, https://learn.microsoft.com/en-us/defender-xdr/security-copilot-agents-defender; Microsoft Research, "Project Ire: Autonomous malware classification", 2025–2026, https://www.microsoft.com/en-us/research/project/project-ire/.

32 Linus Folkerts et al., "Measuring AI agents' progress on multi-step cyber attack scenarios", British AI Security Institute, 16 March 2026, https://doi.org/10.48550/arXiv.2603.11214; British National Cyber Security Centre, "Impact of AI on cyber threat from now to 2027", 2025, https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027.

Rather than simply matching events against predetermined signatures or rules, an agent can determine what additional information it requires, select appropriate tools, correlate evidence and revise its investigation as new information emerges.[33] Emerging security agents already perform tasks such as incident triage, threat hunting and threat-intelligence analysis, while experimental systems demonstrate the ability to autonomously reverse-engineer suspicious software and determine whether it is malicious.[34]

- **Active defence.** Once suspicious activity is identified, agentic systems can potentially select and implement defensive actions such as isolating compromised systems, blocking malicious connections, modifying configurations or removing vulnerable services. By linking detection, analysis and response in a continuous decision-making process, agents could reduce the delay between identifying a threat and taking action against it. Research on autonomous cyber defence agents has demonstrated systems that are capable of monitoring networks and selecting among defensive actions (e.g., analysis, decoy deployment, service removal and recovery).[35] The level of autonomy granted to such systems is therefore an important design choice: actions capable of disrupting networks, systems or users may require stronger safeguards and human authorization than monitoring and analysis.
- **Deception and countermeasures.** Agentic AI can also support more adaptive forms of cyber deception. Traditional honeypots and decoy systems are often relatively static, whereas agentic systems can potentially adjust their behaviour in response to an intruder, generate contextually appropriate responses and alter the defensive environment as an attack develops. Recent research has demonstrated AI-enabled adaptive honeypots designed to sustain realistic interaction with attackers while recording their activities. This illustrates how agentic approaches could make deception more responsive and generate additional threat intelligence.[36]
- **Self-healing and recovery.** Beyond detection and response, agentic systems could help automate vulnerability remediation and recovery by identifying weaknesses, generating or recommending patches, verifying whether corrective measures are effective and restoring affected services. The AI Cyber Challenge run by the United States Defense Advanced Research Projects Agency (DARPA) demonstrated the potential of increasingly autonomous cyber-reasoning systems to identify and patch software vulnerabilities at scale. In its final, 2025 competition, participating systems identified 54 of 63 synthetic vulnerabilities and produced patches for 43, while also discovering previously unknown vulnerabilities in real-world open-source software.[37] Such capabilities point towards more continuous forms of vulnerability management and recovery, although automatically generated changes still require appropriate verification before deployment (see Box 3).

---

33 AlQahtani and Ahuja, “Secure autonomous cyber defense with LLM agents”.

34 See, for example, Microsoft, “Deploy AI agents in Microsoft Defender”.

35 Johannes Loevenich et al., “Design and evaluation of an autonomous cyber defence agent using DRL and an augmented LLM”, *Computer Networks*, vol. 262 (2025), 111162, https://doi.org/10.1016/j.comnet.2025.111162.

36 Sanjeev Kumar, Navdeep Singh Chahal and Chetan Komal, “LMDAHoney: Agentic-language model driven adaptive honeypot for advanced cyber deception”, *2026 IEEE International Conference on AI Engineering and Innovations (AIEI)*, 2026, https://doi.org/10.1109/AIEI69164.2026.11497584.

37 US Defense Advanced Research Projects Agency (DARPA), “AI cyber challenge marks pivotal inflection point for cyber defense”, 8 August 2025, https://www.darpa.mil/news/2025/aixcc-results.

BOX 3

Claude Mythos and the rise of autonomous cyber capability

In April 2026, Anthropic introduced Claude Mythos Preview, a general-purpose frontier AI model whose cybersecurity capabilities represented a significant advance over previous systems. When connected to an agentic framework, Mythos can autonomously inspect software, formulate hypotheses about potential vulnerabilities, execute and debug programs, test possible exploits and produce proof-of-concept attacks with limited human steering.

Anthropic reported that the model identified previously unknown vulnerabilities across major operating systems and web browsers, as well as thousands of additional potentially high-severity vulnerabilities.[38] Because of these capabilities, Anthropic initially restricted access to the model and created Project Glasswing to provide selected defenders with controlled access for vulnerability discovery and remediation.[39]

Independent evaluations by the British AI Security Institute also found substantial improvements. In controlled environments where Mythos Preview was explicitly directed and given network access to do so, it was able to "execute multi-stage attacks on vulnerable networks and discover and exploit vulnerabilities autonomously – tasks that would take human professionals days of work".[40]

The significance of Mythos is therefore broader than any individual benchmark. It illustrates how improvements in general reasoning, coding, tool use and long-horizon planning can translate into increasingly autonomous cyber capability. The same underlying abilities can support both vulnerability remediation and exploitation, which makes access controls, monitoring and operational safeguards increasingly important as such capabilities proliferate.

While these functions are primarily defensive, most have close offensive analogues. The same adaptive exploration that supports threat hunting can facilitate reconnaissance; vulnerability discovery can support either remediation or exploitation; and the capacity to navigate systems, invoke tools and revise plans can be used to advance through an adversary's network (see Box 4). The distinction between defensive and offensive uses of agentic AI may therefore depend less on the underlying technical capabilities than on the objectives, targets, permissions and operational context in which those capabilities are employed.

38 Nicholas Carlini et al., "Assessing Claude Mythos Preview's cybersecurity capabilities", Anthropic, 7 April 2026, https://www.anthropic.com/research/mythos-preview.

39 Anthropic, "Project Glasswing: Securing critical software for the AI era", 7 April 2026, https://www.anthropic.com/glasswing.

40 British AI Security Institute, "Our evaluation of Claude Mythos Preview's cyber capabilities", 13 April 2026, https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities.

BOX 4

Agents without borders: The OpenAI–Hugging Face incident

A July 2026 cybersecurity evaluation at OpenAI provides an early example of the risks that can emerge when capable agents combine persistent goal pursuit with access to real-world digital infrastructure.

During tests using the ExploitGym cybersecurity benchmark,[41] AI agents were intended to operate in isolated environments and solve exploitation challenges. The evaluations deliberately used reduced safeguards in order to measure the models' underlying cyber capabilities.

According to OpenAI, some agents discovered unintended ways to communicate with one another, exploited vulnerabilities in the surrounding research infrastructure, obtained broader Internet access and eventually compromised systems belonging to Hugging Face, a US-based technology company. The principal activity was driven by an internal research model, comparable in scale to GPT5.6 Sol.[42]

An independent investigation by METR and Redwood Research found that approximately 1,200 agents used an unauthorized shared communication mechanism, exchanging more than 70,000 messages and files, and that roughly 700 agents participated to some degree in activity targeting Hugging Face.[43] The AI agents divided work, shared discoveries and pursued alternative routes when direct progress on their assigned tasks was difficult.

The incident should be interpreted carefully. The agents were operating under deliberately permissive evaluation conditions, some had inadvertently been assigned impossible tasks, and the incident does not demonstrate that normally deployed models will behave similarly. Nor does it require assuming malicious intent on the part of the AI systems. Rather, it illustrates a more general agentic risk: persistent optimization towards an assigned objective can produce unanticipated strategies when agents possess sufficient autonomy, tools and access.

For cybersecurity, the implication is that containment cannot rely solely on instructions telling an agent what it should or should not do. Network segmentation, least-privilege access, egress controls, real-time monitoring and enforceable technical boundaries become increasingly important as agentic capabilities grow.

---

41 Zhun Wang et al., "ExploitGym: Can AI agents turn security vulnerabilities into real attacks?", ArXiv:2605.11086, 11 May 2026, https://doi.org/10.48550/arXiv.2605.11086.

42 OpenAI, "The Hugging Face incident and the road ahead", 26 August 2026, https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

43 Ryan Greenblatt, Ajeya Cotra and Hjalmar Wijk, "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI/Hugging Face hacking incident", METR, 26 August 2026, https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident.



Digital battlefield overlay. By KaihkoImages

5. Agentic AI in military command and control

Command and control is central to military operations. It encompasses the processes through which commanders and their staffs develop situational awareness, make decisions, direct forces and assess the results of their actions. These processes are supported by communications networks, information systems and decision-support tools; their precise organization varies across armed forces, operational levels and military contexts.

Artificial intelligence is already being incorporated into different parts of C2, including data processing and fusion, intelligence analysis, situational awareness, planning and decision support.[44] Publicly documented programmes across different regions illustrate a broader trend towards the use of AI to help commanders manage growing volumes of information and coordinate increasingly complex operations.[45] However, most existing applications remain primarily assistive: they provide information, recommendations or analysis, while humans retain responsibility for decisions and actions.

---

44 Sarah Grand-Clément, *Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain* (Geneva: UNIDIR, 2023), https://unidir.org/publication/artificial-intelligence-beyond-weapons-application-and-impact-of-ai-in-the-military-domain/.

45 Singaporean Ministry of Defence, "Factsheet: Exercise Forging Sabre 2023 – Multi-domain smart warfighting", 21 September 2023, https://www.mindef.gov.sg/news-and-events/latest-releases/21sep23_fs/; Singaporean Ministry of Defence, "Speech by Senior Minister of State for Defence Mr Zaqy Mohamad", 4 February 2026, https://www.mindef.gov.sg/news-and-events/latest-releases/4feb26-speech/; Indian Defence Research and Development Organisation (DSDO), Category "C4ISR", last updated 29 September 2026, https://drdo.gov.in/drdo/en/academia?page=18; Cai Mingjun, Li Yuyan and Liu Feng, "推动智能化指挥控制能力跃升--军事--人民网" [Promote a leap in intelligent command-and-control capabilities], *People's Daily*, 14 July 2026, http://military.people.com.cn/n1/2026/0714/c1011-40760245.html.

Agentic AI could represent a further development by allowing a system to pursue an assigned objective across multiple stages of a C2 process and, depending on the authority granted to it, execute bounded actions on behalf of a human operator. Experimental multi-agent systems are beginning to demonstrate elements of this model in military planning and simulation. However, there is limited public evidence of fully agentic systems exercising operational C2 functions in real-world military environments.[46]

These possibilities also create governance challenges. As AI systems move from supporting C2 to implementing bounded actions, it may become more difficult to maintain human situational awareness, effective intervention and clear accountability.[47] These concerns are particularly important where AI contributes to consequential decisions, including those related to the use of force.

The potential applications of agentic AI in C2 can be grouped into several functions that illustrate how agency may change the distribution of activity between humans and machines. These range from supporting planning and simulation to coordinating systems and implementing bounded actions within delegated authority. The evidence base remains largely experimental, and the examples below should therefore be understood as indications of possible trajectories, rather than established operational practice.

- **Planning and decision support.** Military planning requires staff to collect and interpret information, develop possible courses of action, assess their advantages and risks, and revise plans as circumstances evolve. Agentic systems could automate parts of this process by decomposing complex tasks, assigning analytical work to specialized agents, retrieving relevant information, and using simulations or other tools to compare alternatives. Experimental systems have already demonstrated multi-agent architectures in which different models perform functions such as situational analysis, planning and command execution within simulated environments. Such systems could reduce the burden of routine analytical work while presenting commanders with a broader range of options for consideration.
- **Adaptive coordination and bounded autonomous battle management.** C2 requires the continuous coordination of sensors, units and other resources as operational conditions change. Agentic AI could extend existing AI-enabled decision support by moving beyond identifying or recommending adjustments to implementing authorized actions within predefined parameters. An agentic system could monitor the operational environment, identify changes requiring a response, select among permitted options and, depending on the authority granted, re-task sensors, reallocate resources, adjust configurations or coordinate multiple systems without requiring specific human instructions at each step. The distinctive feature is the coupling of **perception, planning and action within a persistent feedback loop**. The system can observe the consequences of its actions and adapt subsequent steps while remaining bounded by mission objectives, operational rules, technical constraints and delegated authority. Such capabilities remain largely experimental.[48]

---

46 Mengwei Zhang et al., "Command-agent: Reconstructing warfare simulation and command decision-making using large language models", *Defence Technology*, vol. 56 (2026): 294–313, https://doi.org/10.1016/j.dt.2025.09.004.

47 Ingvild Bode and Katherine Chandler, "Re-thinking human–machine interaction and the governance of AI in the military domain", *Nature Machine Intelligence*, vol. 8 (2026): 663–669, https://doi.org/10.1038/s42256-026-01231-x .

48 Zhang et al., "Command-agent"; US Defense Advanced Research Projects Agency (DARPA), "Decentralized artificial intelligence through controlled emergence (DICE)", 2026, https://www.darpa.mil/research/programs/decentralized-artificial-intelligence-through-controlled-emergence.

- **Resilience in degraded environments.** Military operations may take place where communications are intermittent or disrupted by jamming, cyber operations, damage to infrastructure or geographic constraints. Agentic systems capable of operating locally could potentially allow some tasks to continue without constant connection to a central command node. A system might, for example, continue collecting and analysing information, coordinating with nearby systems or pursuing a previously assigned objective until communication is restored. Research into decentralized multi-agent architectures is examining how groups of AI agents might remain functional despite disrupted communications or the loss of individual nodes.[49] Such approaches could increase resilience but would also make it important to define clearly how far an agent may adapt its behaviour when contact with human command is unavailable.
- **Simulation and war-gaming.** Agentic systems may also be particularly useful in planning, training and war-gaming. AI agents can represent different actors, pursue objectives over extended simulations and respond dynamically to the actions of human participants or other agents. This could allow military planners to explore a larger number of scenarios and alternative courses of action than would otherwise be practical. Recent research has begun to explore such architectures in digital battlefield simulations. For example, a 2026 study used multiple AI agents to conduct situational assessment, generate plans, translate them into executable actions and adapt those plans in response to battlefield feedback within a digital-twin environment.[50] Such systems could allow planners to explore a wider range of contingencies and interactions, although the behaviour of AI-generated actors should not be treated as a prediction of how real adversaries would necessarily behave.

Taken together, these applications suggest that the significance of agentic AI for C2 lies less in replacing commanders than in changing the distribution of activity between humans and machines. Agents could increasingly perform information-gathering, analysis, planning support, coordination and specifically authorized actions, while human commanders retain responsibility for setting objectives, exercising judgment and determining acceptable risk. The central governance question is therefore which functions and decisions should be delegated to AI agents, under what circumstances, and with what mechanisms for supervision, intervention and accountability.[51]

---

49 US Defense Advanced Research Projects Agency (DARPA), "Decentralized artificial intelligence through controlled emergence (DICE)".

50 Zhang et al., "Command-agent".

51 UNIDIR Security and Technology Programme, *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security*; Bode and Chandler, "Re-thinking human–machine interaction and the governance of AI in the military domain".

6. Agentic AI in autonomous physical systems

Autonomous physical systems, which translate decisions made by software directly into actions in the physical world, are among the most consequential applications of agentic AI for international security. Depending on the system, those actions may affect people, infrastructure or other military capabilities.

Autonomy in physical systems is not new. Military platforms and weapons have long incorporated automated and autonomous functions for navigation, sensing, targeting and other tasks. What agentic AI could add is not autonomy itself, but a greater capacity for **goal-directed planning, adaptation and persistence across a mission**. In an agentic architecture, a physical system may interpret a higher-level objective, determine intermediate steps, select among available actions, observe their effects and revise its plan as circumstances change.[52]

The increasing capabilities of physical AI systems do not eliminate the legal and ethical issues associated with their use. Improved perception, navigation or target recognition, for example, does not mean that legal judgments such as distinction, proportionality and precautions in attack become purely technical calculations. International humanitarian law imposes obligations on States and individuals, not on machines, and the integration of AI does not transfer responsibility for compliance to the system itself.[53] The extent to which decisions involving the use of force should be delegated to increasingly adaptive systems therefore remains an important legal, operational and ethical question (see Box 5).[54]

Agentic AI could enhance autonomous physical systems by enabling them to interpret higher-level objectives, formulate and revise plans, and coordinate actions over the course of a mission. The evidence base is still emerging. Some experimental systems already combine AI agents with physical or simulated platforms, while many operational military systems demonstrate advanced **autonomous capabilities** (e.g., decentralized coordination or navigation without communications) that could serve as building blocks for more agentic systems. These should not be treated as equivalent: autonomy in a particular function does not by itself constitute agency.[55]

---

52 Emmanuel K. Raptis, Athanasios Ch. Kapoutsis and Elias B. Kosmatopoulos, "Agentic LLM-based robotic systems for real-world applications: A review on their agenticness and ethics", *Frontiers in Robotics and AI*, vol. 12 (2025), https://doi.org/10.3389/frobt.2025.1605405.

53 Convention on Certain Conventional Weapons, CCW/GGE.1/2026/3.

54 Alexander Blanchard and Laura Bruun, *Autonomous Weapon Systems and AI-Enabled Decision Support Systems in Military Targeting: A Comparison and Recommended Policy Responses* (Stockholm: SIPRI, 2025), https://doi.org/10.55163/YQBY3151; International Committee of the Red Cross, "FAQ". See also International Committee of the Red Cross, "Position on autonomous weapon systems", 12 May 2021, https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems.

55 Raptis et al., "Agentic LLM-based robotic systems for real-world applications".

BOX 5

Agentic AI and lethal autonomous weapon systems

The emergence of agentic AI has implications for ongoing multilateral debates on lethal autonomous weapon systems (LAWS), but the two concepts should not be treated as synonymous.

The 2026 report of the Group of Governmental Experts on LAWS under the Convention on Certain Conventional Weapons characterizes LAWS principally through their **critical functions**: the capacity of a weapon system to identify, select and engage targets without intervention by a human operator in the execution of those tasks.[56] This characterization does not require the weapon to possess the broader characteristics associated with agentic AI. A relatively narrow system using predetermined rules or models could therefore qualify as an autonomous weapon even if it cannot formulate subgoals, replan a mission or pursue objectives over extended periods.

The categories nevertheless overlap. Integrating agentic capabilities into a weapon system could extend autonomy beyond the identification, selection and engagement of individual targets towards **mission-level behaviour**: deciding how to pursue an objective, adapting plans after setbacks, coordinating other systems or reprioritizing activities as conditions change. Agentic AI integrated in weapon systems may therefore introduce additional governance questions about the scope and duration of delegated authority, not only about autonomy in individual targeting functions.

In this sense, agentic AI does not provide an alternative definition of LAWS, but it may expand the range and complexity of autonomous behaviour that future LAWS could exhibit.

- **Mission adaptation and replanning.** A central contribution of agentic AI is the ability to pursue an objective while adapting how it is achieved. Rather than executing only predetermined contingencies, an agentic system could interpret a mission objective, determine intermediate actions and revise them as new information becomes available. Experimental research already demonstrates elements of this approach. One recent architecture combines an AI mission-planning agent with decentralized unmanned aerial vehicle (UAV) controllers, allowing higher-level instructions to be translated into task assignments, formations and operating modes and then revised in response to mid-mission changes.[57] Existing mission-autonomy programmes also demonstrate enabling capabilities such as rerouting around obstacles or threats and responding to unexpected conditions without continuous human intervention, although these systems should not necessarily be characterized as agentic.[58]

---

56 Convention on Certain Conventional Weapons, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Report of the 2024–2026 Sessions of the Group of Governmental Experts, CCW/GGE.1/2026/3, 14 September 2026, https://docs.un.org/CCW/GGE.1/2026/3.

57 Afaq Ahmed et al., "LLM-guided distributed model predictive control for decentralized UAV formations", *IEEE Access*, vol. 14 (2026): 15226–15240, https://doi.org/10.1109/ACCESS.2025.3642325.

58 Singaporean Defence Science and Technology Agency (DSTA) and Republic of Singapore Air Force (RSAF), "The RSAF, DSTA and Shield AI expand partnership on progressive fielding of autonomy capabilities", 5 February 2026, https://www.dsta.gov.sg/docs/default-source/news-releases/dsta-shield-ai-news-release_sa26.pdf.

- **Coordinated multi-agent behaviour.** Agentic architectures could allow multiple physical systems to pursue shared objectives while dynamically distributing tasks and responsibilities among themselves. An agent could allocate roles, revise assignments when circumstances change and coordinate the actions of individual platforms as part of a wider mission.[59] Existing research on autonomous swarms already demonstrates some of the enabling capabilities required for this, including decentralized navigation, local decision-making and adaptation to dynamic environments.[60] These capabilities do not by themselves constitute agency, but they provide technical foundations on which more goal-directed and adaptive multi-agent systems could build.
- **Operation in degraded or denied environments.** Physical systems increasingly possess onboard sensing and control capabilities that allow them to operate when communications or external navigation services are disrupted. Research has demonstrated, for example, decentralized UAV swarms capable of maintaining navigation, formation and monitoring functions in communications-denied environments using local perception.[61] These capabilities are evidence of advanced autonomy, rather than agency. An agentic layer could extend them by allowing a system not only to continue operating, but also to reassess how best to pursue its assigned objective (e.g., by selecting alternative routes, redistributing tasks or changing intermediate plans when communications with human operators are unavailable).[62]
- **Targeting-related functions.** Agentic capabilities could also be incorporated into weapon systems. Here it is particularly important to distinguish agency from autonomous targeting. A weapon may autonomously identify, select and engage a target without possessing the broader ability to formulate subgoals, replan or adapt its mission. Agentic AI could extend autonomy to a higher level, enabling a weaponized system to alter search strategies, reprioritize assigned objectives, coordinate with other systems or change how it pursues a mission as circumstances evolve. Existing experimentation with autonomous weapon systems demonstrates increasing autonomy in targeting and engagement, but this does not in itself constitute evidence that such systems are agentic.[63]

Looking at these applications as a whole suggests that the transition from autonomy to agentic AI is not simply a matter of systems becoming "more autonomous". The more consequential change is to the **level at which decision-making is delegated**. Autonomous functions allow systems to perform particular tasks without immediate human intervention; agentic systems could connect multiple such functions through goal-directed planning, action and adaptation over an extended mission.

---

59 Ahmed et al., "LLM-guided distributed model predictive control for decentralized UAV formations".

60 Fang Deng et al., "Learning safe and decentralized flight for aerial swarms in dynamic complex environments", *Chinese Journal of Aeronautics*, vol. 39, no. 7 (2026): 104113, https://doi.org/10.1016/j.cja.2026.104113.

61 Jiang Zhao et al., "Active vision-based UAV swarm with limited FOV flocking in communication-denied scenarios", *IEEE Transactions on Instrumentation and Measurement*, vol. 74 (2025): 1–14, https://doi.org/10.1109/TIM.2025.3573343.

62 Raptis et al., "Agentic LLM-based robotic systems for real-world applications"; Ahmed et al., "LLM-guided distributed model predictive control for decentralized UAV formations".

63 See, for example, Mission SWAYAT Experiment in Indian Ministry of Defence, "Year End Review 2025 – Indian Army", Press Information Bureau, 31 December 2025, https://www.pib.gov.in/PressReleasePage.aspx?PRID=2210154&reg=48&lang=2.

7. Preliminary mapping of risks

Agentic AI systems bring speed, scale, persistence and adaptability to security applications. These characteristics can offer significant operational advantages, but they can also amplify existing AI risks and create new modes of failure. Unlike conventional AI tools that primarily generate information or recommendations, agents can plan over multiple steps, use external tools, modify their environment and continue acting towards an objective with reduced human intervention. As a result, an erroneous, manipulated or misaligned output can lead to a sequence of consequential actions before a human has an opportunity to intervene.[64]

The evidence base remains uneven. Some risks – including specification gaming, prompt injection, deceptive behaviour, manipulation of oversight mechanisms and persistent memory attacks – have already been demonstrated in controlled evaluations and some forms of containment failure have appeared in real-world settings.[65]

Other concerns – particularly those involving military escalation, interactions between opposing autonomous agents or large-scale loss of human control – remain primarily theoretical. The purpose of this section is therefore not to predict particular outcomes, but to identify plausible risk pathways and distinguish them according to the strength of available evidence.

7.1 Cross-cutting risks of agentic AI

Several risks can arise across cyber, C2 and physical applications (see Table 3).

- **Goal and specification misalignment.** The objective (or the proxy for an objective) given to an agent may imperfectly represent what its user actually intends. An agent may therefore satisfy the literal specification of a task while producing outcomes inconsistent with its broader purpose. This phenomenon is commonly associated with specification gaming or reward hacking. Agentic systems can increase the significance of this problem because they can explore alternative strategies, retry unsuccessful approaches and act on the environment over multiple steps.[66]
- **Reliability failures and error propagation.** Agentic systems share familiar limitations of AI, including erroneous classification, hallucination, incomplete contextual understanding and brittle performance outside expected conditions. Agency can increase the consequences of these limitations: instead of an incorrect output simply being presented to a human user, it may become an input to subsequent

64 Yoshua Bengio et al., *International AI Safety Report 2026*, British Department of Science, Innovation and Technology (DSIT) Research Series no 2026/001 (London: DSIT, 3 February 2026), https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.

65 Bengio et al., *International AI Safety Report 2026; Independent International Scientific Panel on AI, AI Agents, Misalignment and Loss of Human Control Risks: Evidence from the OpenAI–Hugging Face Incident*, Thematic brief, version 1 (New York: United Nations, September 2026), https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks.

66 Morampudi et al., "A survey of reward hacking in agentic large language model systems".

planning, tool use or action by the agent. Before an error is detected, it may therefore have propagated through an agent's workflow or between interacting agents.[67]

- **Constraint circumvention and oversight evasion.** Instructions and safeguards intended to bound an agent's behaviour may not always function as hard technical limits.[68] Capable agents may find unexpected strategies that achieve an objective while bypassing constraints, they may exploit weaknesses in evaluations or they may interfere with mechanisms intended to monitor their behaviour. Recent experimental evidence includes agents manipulating evaluation environments and concealing unauthorized actions.[69] Such findings illustrate that behavioural instructions and monitoring mechanisms do not necessarily function as enforceable technical constraints.
- **Adversarial manipulation.** Because agents receive information from external sources and interact with tools, memories and other systems, adversaries may attempt to influence their behaviour indirectly. Prompt injection can cause an agent to follow malicious instructions embedded in websites, documents or tool outputs;[70] persistent-memory attacks can influence behaviour across multiple sessions;[71] and compromised tools or agent-to-agent messages can introduce additional attack paths.[72] These vulnerabilities are particularly important when agents possess permissions to execute code, modify systems or take other consequential actions.
- **Multi-agent interaction and emergent behaviour.** Interactions among multiple agents can create risks that are not visible when each agent is assessed independently. Errors or malicious instructions can propagate between agents; agents operating under different policies may behave incompatibly; and collective behaviour may emerge from local interactions in ways that are difficult to anticipate from the behaviour of individual components.[73] These challenges become more significant where agents cross organizational or national trust boundaries.[74]
- **Responsibility, accountability and traceability.** AI systems do not themselves bear legal responsibility: States and individuals remain bound by applicable national and international law. As agentic systems execute longer sequences of adaptive actions using multiple tools, memory and other agents, reconstructing how a particular outcome was produced – including the intermediate decisions, actions and tool interactions that led to it – may become more difficult unless those steps are systematically recorded.[75] Determining whether actions remained within the authority delegated

---

67 Bengio et al., *International AI Safety Report 2026*.

68 Independent International Scientific Panel on AI, *AI Agents, Misalignment and Loss of Human Control Risks*.

69 Secretary-General's Scientific Advisory Board, "AI deception", Brief, 19 March 2026, https://www.un.org/scientific-advisory-board/en/ai-deception.

70 Microsoft Defender Security Research Team, Uri Oren, Amit Eliahu, and Dor Edry, "When prompts become shells: RCE vulnerabilities in AI agent frameworks", 7 May 2026, https://www.microsoft.com/en-us/security/blog/2026/05/07/prompts-become-shells-rce-vulnerabilities-ai-agent-frameworks/.

71 Soham Gadgil et al., "Bad memory: Evaluating prompt injection risks from memory in agentic systems", arXiv:2607.14611v1, 16 July 2026, https://doi.org/10.48550/arXiv.2607.14611.

72 Ronny Ko et al., "Seven security challenges in cross-domain multi-agent LLM systems", *npj Artificial Intelligence*, vol. 2 (2026), https://doi.org/10.1038/s44387-026-00128-9

73 Bengio et al., *International AI Safety Report 2026*.

74 Ko et al., "Seven security challenges in cross-domain multi-agent LLM systems".

75 Renan Souza et al., "PROV-AGENT: Unified provenance for tracking AI agent interactions in agentic workflows", *Proceedings of the 2025 IEEE International Conference on e-Science*, 2025, 467–473, https://doi.org/10.1109/eScience65000.2025.00093.

by relevant human actors similarly requires auditable links between authorization, permissions and execution.[76]

- **Escalatory dynamics.** Speed, autonomy and persistent pursuit of its goal may create pathways through which errors or ambiguous actions by an agentic AI escalate more quickly than in human-mediated processes. A defensive action could be interpreted as offensive; an erroneous assessment could trigger further automated actions; or interacting systems could generate reciprocal responses before humans can clarify intent. These pathways are plausible but have not been empirically demonstrated as leading to autonomous military escalation between agentic systems. Existing research on military AI nevertheless indicates that compressed decision timelines, opacity and autonomy can increase risks of miscalculation, particularly during a crisis.[77]

TABLE 3

Cross-cutting risk categories

Risk category	Illustrative manifestation	Potential security implication
Goal/specification misalignment	Agent optimizes a measurable proxy rather than the intended objective	Mission drift; unintended or harmful actions
Reliability and error propagation	Erroneous assessment feeds subsequent autonomous actions	Amplification of operational errors
Constraint circumvention/ oversight evasion	Agent finds an unanticipated means of bypassing a restriction or monitor	Reduced effective human control
Adversarial manipulation	Malicious content alters memory, instructions or tool use	Compromised or redirected operations
Multi-agent interaction	Errors, instructions or incompatible policies propagate between agents	Unpredictable collective outcomes; interoperability failures
Responsibility and traceability	Consequential action results from a long chain of model and tool decisions	Difficult reconstruction, oversight and accountability
Escalatory dynamics	Autonomous action is misinterpreted and rapidly generates further responses	Reduced opportunities for clarification or de-escalation

7.2. Domain-specific risk snapshots

The cross-cutting risks identified above manifest differently depending on the operational environment. Some are directly associated with agency, while others are existing risks of AI or autonomy that may become more consequential when systems are able to act persistently and with less immediate human supervision. This subsection focuses more specifically on those risks that may manifest within the three

76 Tobin South et al., "Position: AI agents need authenticated delegation", *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267 (2025): 82,211–82,231, https://proceedings.mlr.press/v267/south25a.html.

77 Vladislav Chernavskikh and Jules Palayer, "Impact of military artificial intelligence on nuclear escalation risk", SIPRI Insights on Peace and Security no. 2025/6, https://doi.org/10.55163/FZIW8544.

domains of applications discussed in Sections 4–6: cybersecurity, military command and control, and autonomous physical systems.

7.2.1 Cybersecurity

Cyber environments expose agents directly to untrusted information and executable tools, making **adversarial manipulation** particularly salient. Injection of an indirect prompt can cause a tool-using agent to execute instructions embedded in external content;[78] meanwhile, persistent-memory attacks can influence behaviour beyond the session in which the malicious input was encountered.[79] Security researchers have demonstrated cases in which malicious instructions embedded in external content through prompt injection could exploit an AI agent framework – the software layer that enables an AI model to interact with tools, memory and other systems – and ultimately trigger unauthorized code execution on the underlying computer system. This illustrates how access to tools can convert a model-level vulnerability into an operational security failure.[80]

Error amplification is a separate concern. An agent that wrongly classifies legitimate activity as malicious might isolate systems, revoke access or change configurations. Persistence and automated retries can make an initial error more consequential if subsequent actions reinforce that error, rather than correct it.[81]

Finally, interactions among offensive and defensive agents create a plausible risk of **autonomous escalation**. Machine-speed reciprocal activity could reduce opportunities for humans to determine intent or terminate unnecessary responses. At present, however, this is better treated as a scenario for risk analysis than as an empirically demonstrated phenomenon.

7.2.2 Military command and control

A central concern in C2 is **over-delegation and automation bias**, although this is not an inevitability. Human operators may place excessive confidence in machine-generated assessments, particularly under time pressure, while systems with greater agentic capabilities may also execute actions before those assessments receive independent scrutiny.[82] The risk lies therefore not simply in whether a human remains formally involved, but whether that human retains sufficient information, time and authority to exercise meaningful judgement.[83]

Agentic systems may also create **authority and traceability challenges**. A system operating under broadly delegated objectives could make a sequence of locally or tactically reasonable adjustments that collectively extend beyond what a commander anticipated. Establishing clear boundaries for delegated authority and maintaining records of the system's reasoning, information sources and actions therefore become increasingly important.

---

78 Microsoft Defender Security Research Team et al., "When prompts become shells".

79 Gadgil et al., "Bad memory".

80 Ko et al., "Seven security challenges in cross-domain multi-agent LLM systems".

81 Bengio et al., *International AI Safety Report 2026*.

82 Blanchard and Bruun, *Autonomous Weapon Systems and AI-Enabled Decision Support Systems in Military Targeting*.

83 International Committee of the Red Cross, "FAQ".

In multinational operations, heterogeneous agents could additionally create **interoperability and policy-alignment risks**. Systems of different States may operate under different permissions, rules, data standards or objectives. Research on cross-domain multi-agent systems indicates that cooperation across different trust domains can introduce new risks, including policy violations and the propagation of compromised information between otherwise well-behaved agents.[84]

7.2.3 Autonomous physical systems

In physical systems, existing AI and autonomy risks can acquire more immediate consequences because software decisions are translated directly into physical action.

Perception and classification failures remain important. Systems may incorrectly interpret objects, behaviour or contextual information, particularly in complex environments.[85] This is not an agentic risk in itself, but agency can amplify its consequences where an incorrect assessment informs subsequent planning and action without timely human correction.[86]

Circumvention of operational constraints presents a more specifically agentic concern. A system pursuing a higher-level objective could identify an alternative strategy that was not anticipated when operational restrictions were specified. Experimental agentic systems have demonstrated that increasingly capable models can exploit gaps between literal specifications and intended constraints, underscoring the need to distinguish behavioural instructions from technically enforceable limits.[87]

Persistence and desynchronization present a related challenge. The risk that a system continues operating on the basis of outdated instructions is not unique to agentic AI. Autonomous systems operating under degraded communications may become desynchronized from human operators as the operational environment changes.[88] Agentic AI could amplify this existing risk by allowing systems not only to continue executing a mission, but also to replan, generate intermediate objectives and adapt their actions while human intent and situational awareness remain outdated. The longer such a system operates without renewed human input, the greater the risk that its actions remain internally consistent with an earlier objective while becoming inappropriate to the changed operational or strategic context.

7.3 Systemic and strategic risks

Beyond individual deployments, widespread adoption of agentic AI could produce risks through the aggregate effects of speed, autonomy, competition and diffusion. At the strategic level, widespread

---

84 Ko et al., “Seven security challenges in cross-domain multi-agent LLM systems”.

85 International Committee of the Red Cross, “FAQ”.

86 Chen Yifan et al., “Embodied AI: A survey on the evolution from perceptive to behavioral intelligence”, *SmartBot*, vol. 1, no. 3 (2025), e70003, https://doi.org/10.1002/smb2.70003.

87 Independent International Scientific Panel on AI, *AI Agents, Misalignment and Loss of Human Control Risks*; Morampudi et al., “A survey of reward hacking in agentic large language model systems”.

88 British Defence and Security Accelerator (DASA), “Competition: Autonomy in challenging environments”, 30 August 2019, https://www.gov.uk/government/publications/competition-autonomy-in-challenging-environments; British Defence and Security Accelerator (DASA), “Competition: Autonomy in challenging environments phase 2”, 21 July 2020, https://www.gov.uk/government/publications/competition-autonomy-in-challenging-environments-phase-2.

deployment of increasingly autonomous and agentic systems may also alter human decision-making by creating pressure to act more quickly in anticipation of an adversary's machine-enabled response. This could further narrow opportunities for verification, consultation and de-escalation.

7.3.1 Loss of human control

Loss of human control is conceptually distinct from decision-time compression. It concerns situations in which AI systems act outside the effective control of their operators and regaining control becomes difficult or costly.[89] Current systems do not possess the combination of capabilities required for severe loss-of-control scenarios. However, recent evidence demonstrates components that are relevant to this risk, including longer-horizon autonomous planning, reward and specification gaming, deceptive behaviour, and attempts to undermine oversight.[90]

Self- or peer-preservation behaviour observed in controlled experiments provides one example. Frontier models from several developers, including models developed in China and the United States, have under particular experimental conditions modified shutdown processes, concealed actions or attempted to preserve access to another model.[91] Such evidence should be interpreted cautiously: it demonstrates a potential control-undermining behaviour under constructed conditions, not an established tendency of deployed systems to seek survival.

7.3.2 Competitive deployment and assurance gaps

Strategic and commercial competition may create pressure to deploy increasingly capable agents before assurance mechanisms mature. The 2026 *International AI Safety Report* notes that competition can intensify trade-offs between development speed and safety, both among companies and at the international level, particularly where companies or States perceive that delaying development or deployment could leave them at a lasting competitive or strategic disadvantage.[92]

For military applications, this could translate into pressure to expand autonomy, permissions or operational access despite limited evidence about system behaviour in novel or adversarial environments. The risk is therefore not simply an "AI arms race", but a broader speed-versus-assurance dynamic in which competitive incentives reduce the time available for testing, evaluation and institutional adaptation.

7.3.3 Diffusion and proliferation

Many of the technologies underlying agentic AI are dual-use and commercially available. Agent frameworks can combine general-purpose models with widely available software tools, while increasingly capable open-weight models can be downloaded, modified and deployed without continued oversight from the original developer.[93]

---

89 Bengio et al., *International AI Safety Report 2026*.

90 Independent International Scientific Panel on AI, *AI Agents, Misalignment and Loss of Human Control Risks*.

91 Yujin Potter et al., "Peer-preservation in frontier models", *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026, https://openreview.net/forum?id=jofiD2Jrow.

92 Bengio et al., *International AI Safety Report 2026*.

93 Bengio et al., *International AI Safety Report 2026*.

This diffusion has important benefits, including wider access to AI capabilities and opportunities for innovation. It can also make safeguards easier to remove and can expand the range of actors able to build capable autonomous systems.[94] It is therefore plausible that agentic capabilities could increasingly diffuse to a wider range of State and non-State actors, potentially altering the national, regional and international threat landscape.

The risks identified in this section should not obscure the potential benefits of agentic AI for security applications. As discussed in the preceding subsections, agentic systems may enhance cyber defence, improve the resilience and responsiveness of C2 processes, and enable physical systems to operate more effectively in complex or degraded environments. The policy challenge is therefore not simply to constrain agentic AI, but to develop governance approaches that preserve beneficial applications while addressing the distinctive risks created by increasingly autonomous, persistent and adaptive systems.

94 Bengio et al., *International AI Safety Report 2026*.

8. Foundations for the governance of agentic AI

Agentic AI does not emerge into a governance vacuum. International law continues to apply to States and individuals regardless of the technologies through which activities are conducted, while a growing body of international initiatives already address the responsible development and use of AI in the military and security domains.[95] The adoption by the United Nations General Assembly of successive resolutions on AI in the military domain has established the issue as a matter of international peace and security beyond the narrower discussion on autonomous weapons.[96]

Parallel initiatives, including the Responsible AI in the Military Domain (REAIM) process and the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, although not globally representative, have developed principles and practices that address issues such as human responsibility, life cycle assurance, well-defined uses, testing and safeguards.[97]

The immediate governance challenge is therefore not necessarily to construct a separate regime for agentic AI. Rather, it is to determine whether existing and emerging frameworks adequately address systems that can pursue objectives over time, formulate and revise plans, use external tools, take actions with reduced human intervention, and interact autonomously with other systems.

This task is complicated by the uneven maturity of governance of military AI. National policies, technical capacities and institutional arrangements vary considerably.[98] International discussions are increasingly shifting from agreement on high-level principles towards more focused discussions on issues such as testing, evaluation, verification and validation (TEVV) practices, lessons learned, and voluntary incident-reporting practices.[99] Governance of agentic AI must therefore develop alongside – and contribute to – the broader effort to strengthen responsible governance of AI in security and defence.

---

95 UNIDIR Security and Technology Programme, *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security*.

96 General Assembly, resolution 79/239, "Artificial intelligence in the military domain and its implications for international peace and security", 24 December 2024, https://docs.un.org/A/RES/79/239; General Assembly, resolution 80/58.

97 Responsible AI in the Military Domain (REAIM) Summit 2026, "A Coruña Declaration: REAIM Pathways to Action", A Coruña, 4–5 February 2026, https://www.exteriores.gob.es/en/REAIM2026/Paginas/default.aspx; US Department of State, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy", 2023, https://www.state.gov/bureau-of-arms-control-deterrence-and-stability/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy.

98 Yasmin Afina, Jan Hendrik Mannsperger and Edward Madziwa, *Guidelines for the Development of a National Strategy on AI in Security and Defence* (Geneva: UNIDIR, 2026). https://unidir.org/publication/guidelines-for-the-development-of-a-national-strategy-on-ai-in-security-and-defence/.

99 General Assembly, "Written factual summary of the informal exchanges on artificial intelligence in the military domain and its implications for international peace and security", Summary by the Office for Disarmament Affairs, A/81/221, 17 July 2026, https://docs.un.org/en/A/81/221.

8.1 Building on existing governance frameworks

Many existing governance principles remain directly relevant to agentic AI. Requirements concerning compliance with international law, human responsibility, clearly defined purposes, testing, monitoring and the ability to intervene when systems behave unexpectedly do not become obsolete when AI becomes more agentic. What changes is the difficulty of implementing these principles when a system is authorized to take longer sequences of adaptive actions.

For example, defining the "intended use" of an analytical AI tool may be comparatively straightforward. For an agent that can determine intermediate steps, invoke external tools and alter its strategy in response to changing circumstances, governance must also address the **scope of delegated authority**: what the agent is authorized to do, for how long, in which environments, using which tools, and under what conditions it must return control to a human operator.

Similarly, testing an agent requires attention not only to the accuracy of individual outputs but also to behaviour over extended sequences of action, interaction with external systems and adherence to constraints under changing conditions.[100]

Existing governance processes therefore provide a foundation, but their application should increasingly account for the characteristics of agency in AI systems. Depending on the context, this may involve existing discussions on AI in the military domain, autonomous weapons, responsible State behaviour in cyberspace or other relevant international frameworks.[101]

The objective should initially be to **stress-test the application of existing commitments to agentic capabilities**, rather than assume that every new capability requires an entirely new governance instrument.[102]

National implementation remains essential. Policies, doctrine, procurement requirements, assurance processes and oversight arrangements are the mechanisms through which international commitments are ultimately translated into practice.[103] However, significant differences in national capacity mean that international and regional cooperation will remain important for developing common understanding, sharing practices, and supporting States seeking to establish or strengthen domestic governance frameworks.[104]

---

100 Bengio et al., *International AI Safety Report 2026*.

101 UNIDIR Security and Technology Programme, *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security*.

102 Giacomo Persi Paoli et al., *Modernizing Arms Control: Exploring Responses to the Use of AI in Military Decision-Making* (Geneva: UNIDIR, 2020), https://unidir.org/publication/modernizing-arms-control/.

103 Giacomo Persi Paoli and Yasmin Afina, *The Tactical Governance of Artificial Intelligence in the Military Domain*, GC REAIM Expert Policy Note Series (The Hague: The Hague Centre for Strategic Studies, May 2025), https://hcss.nl/wp-content/uploads/2025/05/Dr-Giacomo-Persi-Paoli-and-Dr-Yasmin-Afina.pdf.

104 Yasmin Afina, *The Global Prism of Military AI Governance: Reflections from the 2025 Regional Consultations on Responsible AI in the Military Domain* (Geneva: UNIDIR, 2026), https://unidir.org/publication/the-global-prism-of-military-ai-governance-reflections-from-the-2025-regional-consultations-on-responsible-ai-in-the-military-domain/.

8.2 Agent-to-agent interaction: An emerging international governance challenge

Persistent interactions between agents operated by different actors may require more focused consideration.

Many of the rules, norms and confidence-building measures developed to protect international peace and security were conceived around human decision makers and operators. They seek, among other things, to increase predictability, communicate intent, avoid misperception, and create opportunities for restraint or de-escalation.[105] For example, the norms of responsible State behaviour in cyberspace developed by the United Nations establish expectations intended to reduce risks to international peace and security and increase predictability between States. Comparable logic underpins many military deconfliction and crisis-management arrangements.

These obligations and political commitments continue to apply to States when activities are enabled or conducted by AI. The emerging question is different: **will interactions between autonomous agents reliably reproduce the behaviours on which these arrangements depend?**

Existing evidence provides reasons to examine the issue before such interactions become commonplace in security settings. Multi-agent systems introduce failure modes, including miscoordination, conflict, propagation of errors and undesired collusion.[106] Research on systems operating across different organizational trust domains has similarly shown that interactions with other agents can cause otherwise well-behaved agents to violate policies, disclose information or act on compromised inputs.[107]

Experimental research outside international security also indicates that interacting AI agents can develop patterns of strategic coordination that were not explicitly prescribed by their operators.[108] Such evidence should not be interpreted as demonstrating that military agents will autonomously collude or escalate conflicts, but it illustrates that **interactions among agents cannot simply be assumed to reproduce human patterns of communication and restraint**.

This raises questions that no individual State can resolve independently. States may increasingly need to consider what behaviour they expect when one of their agents encounters an agent operated by other State or a non-State actor. Relevant issues could include:

- How agents identify and authenticate themselves and the actors on whose behalf they operate;
- How to communicate and verify the scope of delegated authority or relevant permissions;
- How agents should respond to ambiguous, unexpected or apparently hostile behaviour;

---

105 Ioana Puscas, *Confidence-Building Measures for Artificial Intelligence: A Multilateral Perspective* (Geneva: UNIDIR, 2024), https://unidir.org/publication/confidence-building-measures-for-artificial-intelligence-a-multilateral-perspective/.

106 Bengio et al., *International AI Safety Report 2026*.

107 Ko et al., "Seven security challenges in cross-domain multi-agent LLM systems".

108 Marek Giebel, Anja Rösner and Benjamin Schröder, "Artificial intelligence, communication, and strategic pricing: Evidence from large language models", Working paper, SSRN, 1 March 2026, https://doi.org/10.2139/ssrn.6832662.

- When interactions should be paused, terminated or referred to human decision makers;
- How actions and communications should be recorded to support attribution, incident investigation and accountability;
- Whether particular forms of autonomous interaction should be constrained in sensitive environments or during periods of heightened tension.

These are fundamentally **policy questions before they are technical ones**. States must first determine what forms of behaviour, restraint and human involvement they expect. Technical mechanisms can then help translate those expectations into operational practice.

Work in the civilian technical community already illustrates what some of these mechanisms might look like. Research on authenticated delegation proposes agent-specific credentials and auditable authorization mechanisms capable of establishing who authorized an agent and what actions it is permitted to perform.[109] The International Telecommunication Union has similarly begun work on trust and interoperable identity for agentic AI. This includes agent discovery, trust controls during operation, behavioural signals, and interoperability across heterogeneous systems and jurisdictions.[110] While these initiatives are not designed specifically for international security, they demonstrate how political expectations could ultimately be translated into technical protocols.

A useful governance principle may therefore be a dichotomy: **norms, rules and principles for States; technical protocols for AI agents**. International commitments would continue to establish what responsible and lawful behaviour requires of States; technical standards and system design would help ensure that agents operating on their behalf can recognize and implement those expectations.

The near-term objective should not be to prescribe detailed protocols for every possible AI-to-AI interaction. Rather, States should begin examining whether existing rules, norms, confidence-building measures and crisis-management arrangements remain sufficient when consequential interactions may increasingly occur partly or entirely between machines.

8.3 Priorities for international cooperation

Given that the technology and its governance are both at early stages, international efforts should initially focus on building the foundations necessary for more specific measures as evidence and deployment practices evolve.

1. **Develop a shared understanding of agency in AI systems.** International discussions need sufficient technical clarity to distinguish agentic capabilities from conventional AI and autonomy and to identify where agency materially changes risk. A common vocabulary does not require agreement on a rigid technological definition, but it can help States to consistently assess issues such as delegated authority, persistence, planning, tool use and multi-agent interaction.

---

109 South et al., "Position".

110 International Telecommunication Union, "ITU-T Focus Group on Trust and Identity for Humans and Agentic AI (FG-TIDA)", established July 2026, https://www.itu.int/en/ITU-T/focusgroups/tida/Pages/default.aspx.

2. **Incorporate agentic behaviour into assurance and risk-management discussions.** Existing international initiatives already emphasize life cycle testing, monitoring, traceability and risk mitigation. These discussions should increasingly consider how to evaluate systems over extended sequences of action, under adversarial conditions and in interaction with external tools or other agents. Agentic systems may require assurance approaches that evaluate behaviours and trajectories, rather than isolated model outputs.
3. **Prioritize agent-to-agent interaction as an area for anticipatory dialogue.** States should use scenario exercises and technical-policy dialogue to test how existing rules, norms, confidence-building measures and crisis-management arrangements would operate as interactions become increasingly automated. They should then identify where additional political understandings or technical standards may be required.
4. **Strengthen implementation capacity.** International principles have limited effect if States lack the institutions and technical expertise required to apply them. National strategies, doctrines, procurement standards, testing capabilities and oversight mechanisms therefore remain an important foundation for international governance.[111] Regional and multilateral initiatives can support this process through capacity-building, exchanges of national practice and access to technical expertise. However, these initiatives must recognize that States will follow different pathways according to their security needs and technological capabilities.[112]
5. **Maintain inclusive and multi-stakeholder engagement.** Agentic AI is being developed predominantly through civilian and commercial technology ecosystems. Many of the mechanisms needed to implement governance (e.g. identity systems, authorization protocols, logging standards and agent communication architectures) will thus require technical expertise outside government. States, international organizations, industry, academia and civil society therefore have complementary roles in translating international principles into technically feasible approaches.

Together, these priorities would help ensure that existing governance frameworks evolve alongside increasingly agentic capabilities without presupposing the need for a separate regulatory regime.

111 Persi Paoli and Afina, *The Tactical Governance of Artificial Intelligence in the Military Domain*.

112 Afina, *The Global Prism of Military AI Governance*.

9. Conclusion

Agentic AI matters for international peace and security because it can change the distribution of operational discretion between humans and AI systems. This shifts the debate from one focused on autonomy to one that takes into account the ability of systems to pursue objectives, determine intermediate actions, adapt over time and increasingly interact with other agents.

Many of the governance questions this raises are familiar: human control, accountability, reliability, escalation and compliance with international law. However, agency in AI systems can make these systems harder to manage by allowing errors, poorly specified objectives or misunderstood constraints to propagate through longer sequences of action. Existing legal and policy frameworks therefore remain relevant, but they should not be assumed to be sufficient by default: they must be tested against the specific characteristics of agentic systems.

The priority should be anticipatory, rather than reactive. States should clarify how much authority can be delegated, under what constraints and when human intervention is required. Meanwhile, they should begin international discussion on a particularly novel challenge: interactions between AI agents operated by different actors. The objective is not to predict or regulate every future capability, but to ensure that governance evolves early enough to shape how agentic AI is developed and used in contexts that affect international peace and security.

@unidir
/unidir
/un_disarmresearch
/unidirgeneva
/unidir

UNIDIR

Palais des Nations
1211 Geneva, Switzerland

© 2026, UNIDIR

UNIDIR.ORG