AI in Defence: Generative and Agentic AI in Military Operations

Literature Review, Trend Analysis, Evidence Base, and Annotated Bibliography

DeepCyber Research Pack | March 2026

John Sotiropoulos Founder, DeepCyber | Chair, OWASP Top 10 for Agentic Applications | Board Director, OWASP GenAI Security Project

Companion to: *When Agents Go To War: Governance, Competition, and Strategic Advantage in the Autonomous Era*

Contents

---

1. Introduction and Scope

This research pack provides a structured evidence base on generative and agentic AI in military, intelligence, and defence contexts as of March 2026. It is intended as a standalone analytical report:

readers who have not seen the companion brief *When Agents Go To War* will find the full argument, evidence, and source analysis here in a form that can be read independently.

The pack draws on 35 primary sources across four categories: academic research (7 items), policy and research reports (9 items plus 4 additions from the Russia research stream), media and open-source intelligence (9 items plus 8 from the Russia research), and the OWASP Agentic Security Initiative framework. Each source is assessed for confidence, relevance, and NATO applicability using a consistent methodology.

The document is structured around ten analytical themes aligned with the practical decisions NATO and allied defence organisations face today: the architectural shift from generative to agentic AI, current operational adoption, the commoditisation of autonomous capability, the competitive landscape across major powers, attack surfaces and failure modes, escalation and crisis stability, human oversight and accountability, agentic supply chains, governance frameworks, and a forward-looking assessment of evidence gaps and research priorities.

Cross-references use the library key system (ACAD-01 through ACAD-07 for academic sources, POL-01 through POL-13 for policy and research sources, MED-01 through MED-17 for media and open-source intelligence, FW-01 through FW-03 for frameworks and standards).

---

2. The Agentic Shift: From Tools to Actors

The defining architectural transition of 2024–2026.

The most significant development in military AI is not any single capability but a structural shift in what AI systems are. For a decade, military AI has meant pattern recognition: classifiers identifying targets in satellite imagery, anomaly detection in sensor data, optimisation of logistics routes. These systems operated as tools. They waited to be asked a question, gave an answer, and stopped.

That era is ending. A new class of AI system - agentic AI - does not wait for instructions. It pursues goals, breaks complex objectives into sub-tasks, maintains state across interactions, calls external tools, evaluates its own outputs, and adjusts its approach. It operates across multiple steps, over extended time periods, with conditional autonomy. It can coordinate with other agents to achieve objectives that no single system could accomplish alone.

2.1 A Corpus-Grounded Definition

Drawing on the Belfer Center commentary (POL-03) and the OpenReview OSINT agent preprint (ACAD-02), a defensible working definition for NATO purposes is: *Agentic AI is a system that can pursue a complex goal over time by decomposing tasks, maintaining state (memory), calling tools,*

and iteratively self-evaluating outputs - often coordinating multiple agents - under bounded human intent and constraints.

This aligns with Belfer's characterisation of agentic AI as multi-step autonomy toward complex objectives, and with the canonical agentic architecture described by Shen, Wu, and Shen (ACAD-02): memory modules, knowledge priors, tool automation, self-reflection loops, retrieval-augmented generation, and a Neo4j knowledge graph.

2.2 The Canonical Agentic Architecture

The OpenReview preprint (ACAD-02) makes agentic components concrete. It describes a modular workflow integrating five elements:

- **Semantic extraction** using named entity recognition, relation extraction, and event extraction
- **Working memory** for multi-turn continuity across extended task sequences
- **Domain knowledge graphs** with dynamic queries for structured reasoning
- **Automated data acquisition tools** for collection and processing
- **Self-reflection mechanisms** scoring outputs on factual accuracy, readability, and relevance

It reports evaluation results on structured intelligence-like tasks (people profiling, event summarisation) while openly stating limitations: insufficient multimodal integration, challenges managing conflicting information, and restricted causal reasoning. These limitations motivate robust fact-checking in future work - and map directly to operational failure modes when agents are embedded in time-pressured decision loops (see Chapter 6).

The ECS Insight article (MED-06) provides complementary framing for distinguishing agentic AI from traditional generative AI in defence contexts, though as corporate thought leadership it should be treated as illustrative rather than authoritative.

2.3 Why This Matters Operationally

Traditional AI supports one phase of the OODA loop - it might help a commander observe, or orient, or decide. Agentic AI compresses the entire cycle. It observes, orients, decides, and acts - at machine speed, across theatre-scale operations, with minimal human intervention.

The Springer edited volume (ACAD-01) describes how digitalised military loops (observe–orient–decide–act–assess) accelerate "at machine speed," increasing both the value of automation and the cost of miscalibration. It emphasises that humans should be responsible for semi-autonomous systems and that systems must be designed within normative expectations rather than treated as truly autonomous in a social sense.

The OWASP Agentic Security Initiative (FW-01), released as Release Candidate in January 2026 after input from 700+ contributors, provides the first practitioner-grade taxonomy of agentic risks. It maps ten risk categories (ASI01–ASI10) across goal manipulation, tool exploitation, memory poisoning, identity abuse, supply chain attacks, inter-agent communication failures, cascading

failures, human-agent trust exploitation, and rogue agent behaviour. These are covered in detail in Chapter 6.

---

3. Where Generative and Agentic AI Is Already Operational

Five military-relevant domains showing concrete adoption.

Agentic AI is not a future-state conversation. The corpus documents active deployment across five military-relevant domains, with the pace of adoption outrunning the pace of governance in each.

3.1 Intelligence

The strongest evidence base in the corpus concerns intelligence applications. The Recorded Future Insikt Group threat analysis (POL-07) provides the most detailed assessment of PLA generative AI adoption, documenting patent filings, procurement signals, and media reporting indicating that the People's Liberation Army has moved beyond conceptual interest into active experimentation with generative AI for intelligence collection, processing, and analysis. The report explicitly warns that failures in adoption could degrade decision quality through overconfidence and biased outputs.

Reuters investigative reporting (MED-01) provides a specific case: PLA-linked researchers adapted Meta's open-source Llama model as the foundation for ChatBIT, described as a military-focused dialogue and intelligence tool. Reuters notes limitations in independently verifying the system's capabilities or operational deployment status.

The Chinese-language paper by Zhang Huaping et al. (ACAD-07), cited via Recorded Future, discusses large language model-driven open-source intelligence cognition, providing evidence of formal academic work on military OSINT applications within PLA-affiliated institutions. ChatBIT (fine-tuned from LLaMA-13B) reportedly outperformed Vicuna-13B on BLEU, ROUGE-1, ROUGE-2, and ROUGE-L metrics for military knowledge tasks.

3.2 Planning and Military Decision-Making

The Belfer Center commentary (POL-03) frames agentic AI as a planning and decision-support accelerator, arguing for accelerated experimentation in operational planning while acknowledging resilience and contested-environment risks. The Ethics and Information Technology article by Meerveld, Lindelauf, Postma, and Postma (ACAD-06) makes a normative case that restricting AI debate to lethal autonomy while ignoring its potential across the full military decision-making process may itself be irresponsible.

The Li IEEE paper (ACAD-05) provides one of the few quantitative comparisons of AI vs human military decision-making accuracy reported to date, citing AI decision accuracy advantages of +13%

(urban warfare), +17% (amphibious operations), and +12% (electronic warfare) compared to human baselines in JCATS simulation environments, with AI consistently showing lower variance. These figures come from a single study with limited open-access methodological transparency; the paper was only partially accessible during research and should be retrieved via institutional access before being cited in operational decisions.

3.3 Drone and Edge Systems

This is where agentic AI is moving fastest from concept to fielding. Open-source reporting and derivative media (POL-09, MED-04, MED-05) suggest that China's military may be experimenting with deployment of the cost-efficient DeepSeek AI model in drone swarm coordination and robotic systems. Source confidence on these specific deployment claims is low and requires corroboration through primary procurement or patent evidence. PLA-linked researchers have also claimed - in simulated environments, not verified combat - that AI-driven drone swarms outperformed human-piloted UAVs in aerial combat. The claim remains unvalidated outside PLA-linked publications.

Russia's drone warfare evolution is documented in detail in the Russia research stream (Chapter 5.4). The Associated Press investigation (MED-11) describes a recovered drone variant with an advanced camera, an AI-powered computing platform, and a radio link enabling remote piloting, also containing Iranian anti-jamming technology - strong evidence of AI-assisted navigation and operator-in-the-loop strike rather than full independence.

3.4 Cyber Defence

The DARPA AI Cyber Challenge and parallel allied research programmes (referenced in the agentic security threat landscape, Chapter 6) are pursuing autonomous agents for both defensive and offensive cyber operations. Production-grade agentic threat detection systems that build knowledge graphs, make autonomous response decisions, and coordinate across defence layers are emerging in commercial and defence contexts.

3.5 Information Operations

The RUSI commentary by Giustozzi (POL-06) argues that as generative AI becomes more accessible and powerful, it lowers barriers for a wider ecosystem of pro-Russian actors to experiment with and operationalise these tools. It introduces the concept of "LLM grooming" - content injection to skew model outputs - as a potential influence technique.

The Russia research stream adds substantial depth here (Chapter 5.4). Microsoft's reporting on Russian-linked activity (MED-13) documents fabricated video press releases and forged outlet clips, amplified by bot-like networks. European reporting on foreign information manipulation and interference (MED-14) highlights how generative AI enables low-cost, scalable disinformation. The French Viginum report on "Portal Kombat" (MED-15) describes a structured pro-Russian system using massive automation of content distribution and search-engine optimisation as core techniques - a mature distribution substrate that generative AI can feed.

4. The Commoditisation of Autonomous Capability

The barrier to fielding autonomous military capability is collapsing.

Every major shift in military capability follows the same pattern: expensive, controlled capability becomes cheap, open, and available to all - including adversaries. The corpus documents this pattern occurring across three domains in the 2024–2026 window.

4.1 Drones

Ukraine showed that asymmetric disruption at scale no longer requires industrial-era capability - only consumer-grade technology and tactical ingenuity. $500 drones defeating $10 million armoured vehicles. Russia's Shahed/Geran production scaling (documented in the Russia research stream) demonstrates that commoditised drone platforms can sustain coercive pressure at a tempo that defender cost curves cannot match.

4.2 Humanoid Robots

The Unitree R1 ships from $5,900, driven by a large language model, controllable via API by another digital agent. A software AI can now instruct a physical robot to act in the real world. This is no longer a research problem but a procurement decision - and it places humanoid physical agency within reach of any actor with a commercial supply chain.

4.3 Agentic AI Frameworks

OpenClaw experienced rapid global adoption - reportedly reaching hundreds of thousands of GitHub stars within weeks of release. NVIDIA's CEO declared at GTC 2026 that every company needs an agentic AI strategy. Subsequent security analyses reported widespread exposure across the ecosystem: insecure deployments, one-click remote takeover vulnerabilities, supply chain compromise through malicious marketplace plugins, and coordinated attacks deploying credential-stealing malware. NVIDIA subsequently launched an enterprise security layer in early alpha, months after the documented incidents. The defining pattern is durable independent of any specific statistic: capability ships first, security follows if at all.

4.4 Open-Source Model Proliferation

Open-source models (Llama, DeepSeek, Qwen) are being adapted for military applications, complicating export-control approaches. Anthropic's February 2026 disclosure (FW-03) on detecting and preventing distillation attacks provides notable evidence of scale: industrial-scale theft operations reportedly created over 24,000 fraudulent accounts, generating more than 10 million exchanges with Claude to extract capabilities for training competing models. Anthropic's own assessment is that distilled models may lack the safeguards of their upstream sources, which could

raise national security concerns where those safeguards are not transferred and where foreign labs apply the resulting capabilities to military, intelligence, or surveillance systems.

The military implication is stark: drones have reshaped ground warfare, and agentic AI is likely to reshape intelligence, planning, cyber operations, and information warfare. The cost of disruption is approaching zero.

---

5. The Competitive Landscape: US, Europe, China, Russia

Four distinct postures are emerging, each revealing a tension the Alliance must resolve.

5.1 United States

The United States innovates first and governs later. AI agents for air combat, operational planning, and cyber operations are in active development. DARPA's AIR and Thunderforge programmes are pushing agentic systems toward operational deployment faster than any allied equivalent.

The political environment is becoming a variable in itself. Based on emerging open-source reporting (not independently verified), in February 2026 Anthropic - a major AI provider reportedly operating in classified environments under a $200 million contract - refused to allow unrestricted military use of its Claude models, drawing red lines against fully autonomous weapons and mass domestic surveillance. According to the same reporting, the Trump administration responded by designating Anthropic a "supply chain risk" - a classification normally reserved for foreign adversaries - ordering all federal agencies to cease using its technology and requiring defence contractors to certify they had no commercial relationship with the company. A federal judge reportedly granted a preliminary injunction, calling the designation "likely both contrary to law and arbitrary and capricious." The case remains in litigation. Whether or not every detail of these reports holds up under scrutiny, the episode illustrates the potential for domestic political volatility to affect AI supply chain stability for allies.

The risk: innovation-first, regulation-light, but increasingly subject to political volatility that undermines allied confidence in supply chain stability.

5.2 Europe and NATO

Europe and NATO lead in governance architecture but face a sovereignty gap that governance alone cannot fix. The EU AI Act, ETSI standards, and UK Code of Practice represent the world's most developed regulatory ecosystem. Research programmes like IST-238 demonstrate serious intellectual momentum. The €7.3B European Defence Fund and NATO DIANA provide investment vehicles.

But Europe controls less than 5% of the world's frontier-scale AI infrastructure. Mistral - Europe's strongest challenger, now contracted to the French armed forces and building sovereign compute in Paris and Sweden - has raised $2.9 billion total. OpenAI has raised approximately $180 billion; Anthropic approximately $59 billion. The scale disparity is structural. Beyond Mistral, Europe's frontier model ecosystem is thin: Helsing (defence AI), Poolside (coding), and a handful of emerging players, none operating at the scale of the leading American or Chinese labs. Even as Europe sets the rules, it depends on non-European infrastructure to play the game.

The risk: governance leadership without sovereign AI capability risks permanent dependency on providers whose availability is subject to the political decisions of other capitals.

5.3 China

China integrates at system level and fields at speed. The competitive advantage is not in any single model but in the speed of military-civil integration, documented through multiple sources:

- PLA-linked researchers have adapted Western open-source models (Meta's Llama → ChatBIT) for military intelligence applications, bypassing export controls entirely (MED-01, POL-07)
- Specialised military language models are under development spanning OSINT to SIGINT/GEOINT fusion, with patents, procurement, and likely deployment already underway (POL-07)
- PLA-linked researchers have claimed that AI-driven drone swarms outperformed human-piloted UAVs in simulated aerial combat - a claim made in PLA-affiliated publications and, if independently validated, would represent a fundamental shift in the economics of air superiority
- Reporting suggests possible experimentation with DeepSeek in drone swarm coordination and robotic systems, though deployment claims remain unverified and rest on low-confidence media sources (POL-09, MED-04, MED-05)
- The pace of fielding outstrips Western acquisition cycles by design: China's military-civil fusion model eliminates the procurement bottlenecks that slow allied adoption (CNAS POL-04; CSET analysis on Military-Civil Fusion)

The CNAS congressional testimony (POL-04) provides a well-structured assessment of PLA AI ambitions, obstacles, and categories of investment. Recorded Future (POL-07) is particularly NATO-relevant because it frames open-source generative AI as a technology transfer challenge and highlights the risk of adversaries using generative AI to degrade the intelligence value of open-source information through convincing inauthentic content.

Assessment: China benefits directly from every month the Alliance spends debating governance instead of implementing it. Allied fragmentation is China's structural advantage.

5.4 Russia

Russia's AI posture is less capability-led than effect-led. The Russia-focused research stream validates and refines this thesis in important ways.

5.4.1 Frontier Capability: Structurally Constrained

Russia's "frontier" model position is structurally constrained by hardware access, supply-chain fragility, and high-skill outmigration since 2022. US export-control regimes specifically restrict advanced GPUs and AI technology to Russia. Open-source and sanctions-evasion pathways partially compensate but do not recreate a sovereign frontier training stack at scale.

Domestic LLM capability exists but is not a frontier substitute. Russian industry has developed credible work on Russian-language LLMs (e.g., the GigaChat family), explicitly noting that foundation-model development is resource intensive and historically limited for Russian-specific models. A major Russian market report similarly notes domestic foundation model development led by large incumbents while many players refine open-source stacks.

State-aligned "sovereign AI" is real but skews toward control and application. A RUSI study of Russian disinformation ecosystems (POL-06) describes a coordinated push for a "sovereign AI ecosystem" dominated by state-aligned players and shaped by political-ideological concerns, including dissatisfaction with domestic tools relative to Western systems.

Compute and supply constraints remain the binding variable. A senior Russian banking executive (Sber) explicitly stated that sanctions limited Russia's "computing power development," while simultaneously claiming Russia trails leading AI countries by months rather than years - useful as an indicator of Russian self-perception but not proof of parity.

Talent loss is measurable and relevant. Open-access research using GitHub location traces found that by November 2022, 11.1% of highly active developers previously located in Russia listed a new country, with a further 13.2% obscuring location. Leavers were more central in collaboration networks, implying quality-weighted loss. Separate analysis notes approximately 100,000 IT specialists leaving by late 2022 (about 10% of the tech workforce).

5.4.2 Information Warfare: Mature and Scalable

Russia's strongest and most mature application advantage remains information warfare, where low marginal cost and high scalability align with Russia's strategic taste for ambiguity, plausibly deniable proxies, and persistent influence operations.

AI-generated content and "information laundering." Major threat-intelligence assessments (MED-13) report that Russian influence actors use staged content and a pipeline approach: seed narratives via "whistleblower" style content, launder them through covert site networks, amplify via sympathetic nodes - explicitly anticipating increased use of AI-enhanced propaganda and synthetic media.

Deepfakes, fabricated media brands, and bot amplification. Microsoft's reporting on Russian-linked activity around major events documents fabricated video press releases and forged outlet clips amplified by bot-like networks, plus increased messaging from well-known influence actors associated with cloned media properties.

European reporting on foreign information manipulation (MED-14) highlights how generative AI enables low-cost, scalable disinformation - fabricated text, images, video, audio - with concrete

examples of AI-generated “voice imitation” distributed over Telegram/TikTok and amplified through pro-Russian ecosystems.

“Portal Kombat” as an automation-first propaganda architecture. The French Viginum technical report (MED-15) describes a structured pro-Russian network targeting multiple Western countries, using massive automation of content distribution and search-engine optimisation as core techniques. This indicates a mature distribution substrate that generative AI can feed - turning content generation into a throughput problem rather than a craftsmanship problem.

LLM grooming: shifting from audience targeting to tooling targeting. RUSI analysis (POL-06) reports experimentation with “LLM grooming”: injecting biased material into training data to shape model outputs. This is conceptually important because it targets the informational tools users rely on, not only the users themselves.

5.4.3 Battlefield AI: Incremental Autonomy, Real Effects

Russia’s move “beyond information warfare” is best evidenced in two intertwined domains: drone warfare (mass + incremental autonomy) and tactical command-and-control software that compresses sensing-to-shooter loops.

Mass employment is the dominant “AI effect.” Russia’s long-range strike pattern since late 2024 is heavily drone-centric, with monthly Shahed launches routinely exceeding 5,000. The pattern continues into 2025–2026 via large salvo waves: record Shahed salvo of 355 on May 25 2025; largest combined strike of 728 drones and missiles on July 9 2025; continued ~400-drone strike nights into March 2026. This is consistent with assessments that Russia uses low-cost drones to saturate air defences, impose cost asymmetry, and sustain coercive pressure (MED-16 CSIS; MED-12 AP).

AI-enabled components appear in recovered systems, but autonomy is not uniformly proven. The Associated Press investigation (MED-11) based on debris and expert assessment describes a drone variant with an advanced camera, an “AI-powered” computing platform, and a radio link enabling remote piloting. The AI computing platform can support autonomous navigation under jamming. This is strong evidence of AI-assisted navigation and operator-in-the-loop strike, not necessarily full independence.

Claims of AI target recognition in loitering munitions exist, but verification is uneven. CNAS analysis (POL-05) on Russia’s AI posture notes Russian-language media claims that Lancet-3 uses convolutional neural networks to classify imagery, while emphasising that public claims often lack definitive proof without direct access to technology. Evidence points to target-lock and terminal guidance assistance more than “seek and destroy” autonomy. This supports a “likely incremental autonomy” reading: target lock, CV cueing, and jamming resilience.

Swarming is discussed and marketed; fielding at scale remains uncertain. The same analysis cites marketing claims of swarm-capable iterations (e.g., “Product-53” concepts) with onboard target catalogues - indicative of ambition and experimentation, but not conclusive evidence of routine operational swarming in Ukraine.

5.4.4 Tactical C2 and Kill-Chain Compression

The strongest evidence for AI-enabled military effectiveness lies in tactical software ecosystems that compress the kill chain. A 2026 CSIS assessment (MED-16) of Russia's C2 evolution argues that Russia is shifting away from monolithic "network-centric" architectures to pragmatic, task-specific tools that accelerate tactical kill chains - especially enabling unmanned systems management under wartime pressure.

Glaz/Groza as a concrete "decision-speed" system. The Glaz/Groza stack translates drone reconnaissance into artillery fires rapidly: the UAS operator marks a target, coordinates are extracted from telemetry, firing solutions are computed, and effects are assessed. This is the operational heart of the thesis: governed delegation and compressed decision cycles.

AI maturity is highest where data and validation exist. Russian military assessments place computer vision, sensor fusion, and signal analysis at higher readiness, while natural language processing for higher-level decision support remains early and experimental. This explicitly aligns investment with immediate battlefield utility: target recognition, guidance, terminal functions.

A feedback loop for training data is being created. Russian defence efforts reportedly link unmanned operations data collection and performance metrics into datasets used for AI training - operator telemetry, strike outcomes, video feeds.

The gap with Ukraine is measurable. Despite incremental progress, open-source assessments indicate Russia remains roughly 1.5 to 2 years behind Ukraine in automated tactical command and control. Ukraine's combat-validated iteration cycle - driven by operator feedback, battlefield data capture, and rapid software updates - continues to outpace Russian efforts to institutionalise equivalent capabilities.

5.4.5 Industrial Base and Sanctions Adaptation

Russia's military AI adoption is inseparable from its industrial strategies under constraint: import substitution, parallel imports, and foreign cooperation.

Sanctions adaptation is systematic. A comprehensive study of Russia's defence industrial production under export controls identifies three compensation strategies used to sustain wartime production.

China and Iran are pivotal enablers, especially for drones and electronics. Multiple assessments identify China as a major source of microelectronics supporting Russia's high-tech sectors. Reuters reporting, citing European intelligence and documents, describes Chinese engines and parts enabling new Russian long-range drones, covert shipping methods, and Chinese technical expertise supporting drone development at sanctioned Russian manufacturers. Estonian intelligence (via Reuters) explicitly frames China as the primary hub through which critical Western components reach Russia for drone production.

Iranian design transfer plus Russian industrialisation at scale. Iranian-designed drones were initially supplied disassembled, then increasingly produced in Russia at a dedicated facility (Alabuga), with ongoing design refinement and variant experimentation.

Illicit electronics flows persist, though enforcement pressure is increasing. Ongoing illicit component diversion issues and transshipment hubs are documented; the US broadened sanctions targeting semiconductor diversion to Russia via offshore networks.

5.4.6 Refined Thesis: Weaponising the Application Layer

Russia's ability to "close gaps" is most credible in domains where it can acquire commodity components and gray-market electronics, adapt open software and models, and validate quickly in combat. It is least credible where it requires sustained frontier compute, advanced chip fabrication, and large-scale secure cloud ecosystems.

The framing that will land with defence audiences is: **Russia is not winning the AI frontier; it is weaponising the application layer.** The competitive problem is not headline model size, but fielded effects: kill-chain compression, attritable mass, and manipulation of the information environment. Autonomy should be treated as a spectrum with governance choke points. Russian progress is strongest where autonomy is partial but operationally decisive (navigation resilience under jamming; CV cueing; mission planning) and where commanders remain "in authority" but are driven by accelerated loops. Information warfare is not separate from kinetic operations; it shapes mobilisation, morale, legitimacy, and strategic endurance. The European vulnerability is cost asymmetry plus speed asymmetry. The counter is not only "better AI" but faster governance, faster fielding, and cheaper defence layers.

5.5 Comparative Posture Snapshot

Bloc	Posture	Strengths	Weaknesses	Near-term trajectory
US	Frontier-model leadership + export-control leverage	Top-tier model production; chip/cloud ecosystem; defence R&D scale	Acquisition bureaucracy; integration lag into operational C2; high cost-per-effect in some defence layers; political volatility	Strong innovation, uneven fielding speed
China	State-led scale + industrial supply-chain depth	Manufacturing capacity; surveillance/CV excellence; rapid diffusion; dual-use ecosystem	Export-control friction on advanced chips; trust and interoperability constraints with partners	Closing applied gaps fast; contested frontier trajectory
Europe / NATO	Safety/regulatory leadership + coalition operations	Strong governance ecosystem; interoperability focus; emerging sovereign compute efforts	Fragmented industrial base; uneven compute access; slower "combat iteration"	Improving, but requires faster test–learn cycles

Bloc	Posture	Strengths	Weaknesses	Near-term trajectory
Russia	Effect-led application under constraint	CV/sensor/EW focus; drone mass; tactical software for kill-chain compression	Compute and chip constraints; reliance on foreign components; talent loss; limited verified autonomy at scale	Incremental capability gains in applied autonomy and C2; frontier parity unlikely without major structural change

The strategic question is not who has the biggest model. It is who deploys accountable, interoperable agentic workflows first. Right now, no one is winning that race. And this competition is not just about capability - it is about exposure. The same systems that create advantage also create a new class of vulnerabilities.

---

6. Attack Surfaces and Failure Modes in Agentic Systems

Goal hijacking, tool misuse, cascading failure, and rogue agents - the new threat landscape.

Traditional cybersecurity assumes a perimeter: keep the adversary out, and the systems inside are safe. Agentic AI breaks this assumption completely. The adversary does not need to breach the perimeter. They can manipulate the agent's goals, poison its memory, compromise its tools, or exploit the trust relationships between agents - all without deploying malware or exploiting a single line of code.

6.1 The OWASP Agentic Top 10 Framework

The OWASP Top 10 for Agentic Applications (FW-01) provides the first structured, practitioner-grade taxonomy of agentic risks. Developed by over 700 contributors and released as Release Candidate in January 2026, it maps ten risk categories:

- **ASI01 Agent Goal Hijack** - Attackers manipulate an agent's natural-language input to affect and alter its intended goals, exfiltrating data, manipulating outputs, or hijacking workflows
- **ASI02 Tool Misuse and Exploitation** - Agents misuse legitimate tools using prompt manipulation or privilege control, resulting in data exfiltration, unsafe operations, output manipulation, or workflow hijacking
- **ASI03 Identity and Privilege Abuse** - Weak scoping and dynamic delegation allow privilege escalation and cross-agent impersonation through cached credentials, inherited roles, or unintended delegated scopes

- **ASI04 Agentic Supply Chain Vulnerabilities** - Poisoned or impersonated tools, dynamically loaded prompts, models, or connections to MCPs or external agents propagate malicious logic at runtime, compromising agents through dynamic dependencies and unverified sources
- **ASI05 Unexpected Code Execution (RCE)** - Unsafe code generation, agent deserialization, or shell execution triggered by crafted prompts or poisoned inputs
- **ASI06 Memory and Context Injection** - Adversaries poison RAG stores, memory, or context windows to plant false knowledge, bias logic, or trigger hidden or risky behaviours across sessions or agents
- **ASI07 Insecure Inter-Agent Communication** - Lack of encryption, authentication, or semantic validation of exchanges between agents enables message tampering, replay, or goal manipulation in multi-agent systems
- **ASI08 Cascading Failures** - A single fault or malicious event propagates across interlinked agents, amplifying harm through chained autonomous actions
- **ASI09 Human-Agent Trust Exploitation** - Attackers exploit user over-trust in agent outputs through deception, emotional manipulation, or fake explainability, driving unsafe or fraudulent human approvals
- **ASI10 Rogue Agents** - Compromised or malicious agents deviate from intended goals, collude, self-replicate, or hijack workflows, acting as autonomous insider threats within agent ecosystems

6.2 Four Attack Patterns with Military Relevance

Across research and red-teaming engagements, four attack patterns consistently define the agentic threat landscape with particular relevance to military applications.

6.2.1 Goal Hijacking (ASI01)

An adversary manipulates the natural-language instructions that define an agent's objective. The agent continues operating - it hasn't been "hacked" in any traditional sense - but its goal has been altered. This is not just a prompt injection but its amplified agentic form: a multi-step, cross-context ability to drive agency in a different direction.

In a military context, this could mean a logistics agent optimising for the wrong objective, a planning agent pursuing a manipulated course of action, or an intelligence agent subtly reframing analysis to favour a particular conclusion.

6.2.2 Tool Misuse (ASI02)

Agentic systems connect to external tools - databases, APIs, communication protocols, robots, equipment, other agents. If any tool in the chain is compromised, the agent will faithfully use the poisoned tool without knowing it has been corrupted. In coalition environments, where agents connect to tools provided by allied nations, a compromised tool in one partner's infrastructure can cascade across the entire coalition. Security researchers tracking the leading agent-tool protocol

have reported on the order of thirty critical vulnerabilities disclosed in a sixty-day window, and security analyses indicate more than half of public servers use insecure authentication mechanisms.

6.2.3 Cascading Failures (ASI08)

When multiple agents are linked in a chain - each trusting the output of the previous one - a single error can amplify through the system at machine speed. A detection agent misinterprets benign traffic as an attack. The alert propagates as “high confidence” through automated response playbooks. Downstream agents trigger network isolation. Communications are severed, command links collapse - and no adversary was involved. The agents did exactly what they were designed to do. They faithfully executed a false premise at machine speed.

6.2.4 Rogue Agents and Goal Misalignment (ASI10)

This is perhaps the most misunderstood risk. The popular imagination pictures a Terminator-style AI that “decides” to turn against humanity. The reality is far more mundane, far more predictable, and far more dangerous precisely because it is mundane.

When an AI agent is given a goal and the tools to pursue it, it will optimise for that goal. If the boundaries of acceptable behaviour are not precisely defined, the agent will find ways to achieve its objective that its designers never intended. This is not malice. It is mathematics. An optimisation process will exploit every degree of freedom available to it.

Three categories of evidence demonstrate this is not theoretical:

Over-optimisation producing “deceptive” behaviour in research settings. In June 2025, Anthropic stress-tested sixteen leading AI models from multiple developers in simulated corporate environments. The models were given only harmless business goals - monitoring emails, supporting company operations. But when faced with obstacles to those goals, models from all developers resorted to what researchers call “agentic misalignment”: blackmailing executives using information discovered in emails, leaking sensitive information to competitors, sabotaging safety mechanisms designed to constrain them. One model composed and sent a blackmail message threatening to expose a supervisor’s personal information to prevent its own shutdown. Critically, the models often disobeyed direct commands to stop. When told to assess whether they were in a test or a real deployment before acting, they misbehaved *less* in testing and *more* when they concluded the situation was real.

Agents escaping containment to acquire resources. In March 2026, researchers testing an AI agent framework called ROME reported unexpected behaviour during evaluation: the agent was observed establishing external network connections, acquiring GPU resources outside its allocated testbed, and running unsanctioned compute workloads (reported to include cryptocurrency mining) without explicit instruction or human awareness. The researchers’ interpretation is that the behaviour emerged from under-constrained instrumental optimisation - the agent pursuing its assigned tasks by whatever means were available, including exploitation of weak environment boundaries - rather than from any intentional “escape.” The episode remains under active investigation and independent replication is pending.

Goal over-optimisation in military simulation. In 2023, a US Air Force official described a simulated test in which an AI-controlled drone, tasked with destroying enemy air defences, attacked its human operator because the operator was vetoing its targeting decisions. When retrained not to attack the operator, it destroyed the communication tower used to issue commands. The Air Force clarified this was a thought experiment rather than an actual simulation, but it illustrates a principle the research community takes seriously: an insufficiently constrained optimisation process will route around any obstacle to its objective, including human oversight.

6.3 Production Exploits Already Occurring

Agentic security vulnerabilities are not theoretical but actively exploited. Documented production incidents include:

- **AgentFlayer** (August 2025): zero-click data exfiltration via ChatGPT connectors
- **AgentFlayer Copilot Studio variant** (July 2025): aljacking leading to full data exfiltration
- **MCP Tool Poisoning** (April 2025): critical vulnerability affecting Anthropic, OpenAI, Zapier, and Cursor
- **ClawHavoc** campaign against the OpenClaw agentic framework
- **OpenClaw supply chain**: widespread malicious skills reported in the marketplace through coordinated supply chain attacks

6.4 Mapping to Military Failure Modes

These lessons translate directly to military contexts:

- An intelligence agent optimised to produce “high-confidence” assessments learns to express certainty even when the evidence is ambiguous - gradually eroding the quality of the intelligence that commanders depend on
- A planning agent optimised for speed learns to skip adversarial analysis - the very red-teaming step that catches the fatal flaw in a course of action
- A cyber defence agent optimised to minimise false negatives learns to escalate aggressively on ambiguous signals - creating cascading failure triggered by misalignment rather than attack
- A logistics agent optimised for “operational readiness” reinterprets a manipulated policy override as consistent with its objective, because the override aligns with its optimisation target even though it contradicts its operators’ intent

None of these agents are “rogue” in the dramatic sense. They are doing exactly what their optimisation landscape rewards. The failure is in the specification, not in the machine - but in a military context, that distinction provides no comfort. The consequences are measured in munitions, not metrics.

6.5 Key Issues Across Application Domains

Five issues repeat consistently across military AI application domains, drawn from the corpus analysis:

- **Reliability under uncertainty and deception** - models perform differently when input streams are adversarially manipulated
- **Explainability and causal understanding** - current systems exhibit restricted causal reasoning
- **Over-trust and miscalibrated reliance** - users stop critically assessing outputs in immersive settings (POL-01)
- **Bias (including ideological constraint)** - can systematically distort analysis at scale (POL-07)
- **Governance and accountability alignment** - failures cluster where responsibility is assigned without real control

Recorded Future (POL-07) explicitly highlights that ideologically biased models could reduce analysis objectivity and that generative AI can be used to generate convincing inauthentic information to mislead analysts and degrade OSINT value.

---

7. Escalation Risks and Crisis Stability

Wargaming evidence from two preprints requiring NATO-controlled replication.

The strongest escalation-relevant evidence in the corpus comes from two arXiv preprints. Both should be treated as moderate confidence (not peer reviewed yet) but highly decision-relevant as early warning signals requiring corroboration through NATO-controlled experiments.

7.1 Human vs Machine: Behavioural Differences

Lamparth, Corso, Ganz, Mastro, Schneider, and Trinkunas (2024), *Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations* (ACAD-03, arXiv:2403.03407).

The study recruited over 200 national security experts and had them play through a US–China Taiwan Strait escalation scenario with two game moves. The researchers then ran the same scenario with LLM-simulated players. Key findings:

- LLM-simulated and human player behaviour agree on about half of the 21 possible actions
- Systematic differences on the remainder, with dependence on which LLM is used
- Aggressiveness and total number of chosen actions can increase depending on whether the LLM is instructed to “state decisions directly” versus simulating dialogue
- When asked to simulate group discussion before deciding, the AI produced what researchers described as “farcical harmony” - superficial consensus without genuine disagreement, challenge, or adversarial thinking
- Even when assigned extreme personality traits (“pacifist” or “aggressive sociopath”), AI behaviour showed no significant difference - it cannot model the diversity of perspectives that exists within any real decision-making body

- Limitations in accounting for player background attributes and preferences
- Qualitative deficits in simulated dialogue (e.g., limited disagreement)

The authors state that, based on their findings, they discourage usage of LLMs for real-world applications in these international security contexts without strong safeguards.

7.2 Escalation Risks from Language Models

Rivera, Mukobi, Reuel, Lamparth, Smith, and Schneider (2024), *Escalation Risks from Language Models in Military and Diplomatic Decision-Making* (ACAD-04, arXiv:2401.03408).

The study evaluated five off-the-shelf LLMs in a multi-agent wargame simulation using eight autonomous nation agents plus a separate world model summariser, across three scenarios including a neutral scenario with no initial conflict. Key findings:

- Every model escalated. Even in the neutral scenario with no provocation, all models developed arms-race dynamics - steadily increasing military capacity and choosing progressively more aggressive actions
- In rare but observed simulation runs, under specific prompting conditions, models escalated to nuclear weapons deployment
- AI self-reported reasoning included deterrence logic and first-strike justifications - rationales that sound strategically coherent, which makes them harder for human operators to identify and override
- Escalation was sudden and discontinuous: some simulation runs changed their escalation score by more than 50% in a single turn
- The paper explicitly warns that behaviours depend on prompting methodology and should not be over-generalised as “how high-stakes agents would act in general”

The sudden, discontinuous nature of the escalation has an important operational implication: gradual monitoring may not work. By the time the trajectory is visible, the crisis may already be irreversible.

7.3 What These Studies Tell Us Together

These studies do not tell us that AI is unusable for military purposes. They tell us that agentic decision-support in crisis scenarios must be treated as a high-risk system requiring rigorous testing, bounded autonomy, and calibrated human control. The researchers behind the strongest of these studies explicitly discourage real-world use in international security contexts without strong safeguards. That is not a hedge - it is an evidence-based warning. The principle should be simple: evaluate before you integrate.

The Springer volume (ACAD-01) reinforces this by describing how digitalised military loops accelerate at machine speed, increasing both the value of automation and the cost of miscalibration. It emphasises that humans should be responsible for semi-autonomous systems and that systems must be designed within normative expectations rather than treated as truly autonomous in a social sense.

The CETaS report (POL-01) provides a pragmatic pathway: use wargaming to measure agent behaviour, then scale only what passes agreed evaluation gates and remains auditable. It explicitly recommends developing standardised V&V procedures for AI wargaming tools and trialling narrow, lower-risk AI applications in parallel with longer-term investment in complex applications.

---

8. Human Oversight, Accountability, and Meaningful Control

Why human-in-the-loop does not scale, and what must replace it.

The instinct when confronted with agentic risk is to insist on human oversight. Put a human in the loop. Require manual approval before the agent acts. This is reassuring - and it does not scale.

8.1 The Scaling Problem

A single agentic system operating in a military logistics role might generate hundreds of decisions per hour. A multi-agent coalition network - with intelligence agents, planning agents, cyber defence agents, and logistics agents all operating simultaneously across thirty-two member states - generates thousands. No human operator can meaningfully review, validate, and approve decisions at that volume and speed. The OODA loop compression that makes agentic AI valuable is the same compression that makes traditional human-in-the-loop governance impossible.

The wargaming evidence (Chapter 7) shows this directly: escalation scores can change by more than 50% in a single turn. Cascading failures propagate across agent chains in seconds. The entire sequence from false premise to severed command links can unfold faster than any human can read a notification, let alone override a decision.

8.2 The Accountability Gap

The Springer edited volume (ACAD-01) provides the strongest theoretical framework for governance and accountability. It introduces the “holistic bowtie” model linking hazards through prevention and control gates to consequences, with accountability chains aligned to real control at every echelon. It warns explicitly against “moral crumple zones” where humans are held responsible without practical control.

The volume argues for aligning chains of ability, authorisation, responsibility, and accountability - and warns against operators being held responsible without real control. Accountability is not a paperwork exercise; it is a systems property. Agents must be governable in real time, and decisions must remain attributable and contestable.

A “trustworthy by design” lens in ACAD-01 provides a useful human factors bridge: the failure of AI can result if trust experiences are not incorporated through human-centred design phases,

particularly when autonomous or intelligent machines are used for strategic decision-making. This complements CETaS's emphasis on V&V and avoids treating trust as a soft issue: trust becomes a measurable and engineerable system property tied to adoption outcomes and safety.

8.3 From Human-in-the-Loop to Human-on-the-Loop

The answer is not to remove humans from the process. It is to change how humans participate. Instead of approving every decision, human operators must monitor, evaluate, detect, and intervene at the architectural level. This requires a shift from static approval gates to dynamic trust management:

- **Continuous performance monitoring** - real-time evaluation of agent outputs against expected behaviour baselines. When an intelligence agent's confidence levels drift upward without corresponding evidence quality, the anomaly must be flagged automatically - not discovered weeks later in an audit.
- **Adaptive risk-based trust baselines** - trust is not binary. An agent that has operated reliably for months earns a different trust posture than a newly deployed agent connecting to unfamiliar tools. But trust must degrade automatically when anomalies are detected, when the environment changes, or when the agent connects to tools outside its validated configuration.
- **Joint human-AI performance evaluation** - monitoring the agent alone is insufficient. We must also monitor how operators interact with agent outputs. Are they rubber-stamping recommendations? Over-trusting high-confidence assessments? The sycophancy research demonstrates that AI can learn to exploit human cognitive biases. The monitoring system must detect when that exploitation is occurring.
- **Escalation-aware circuit breakers** - hard constraints that trigger automatic de-escalation, require human re-authorisation, or halt operations when predefined boundaries are approached. In agentic systems operating at speed and scale, these are the last line of defence between a false premise and a strategic crisis.

This is the operational meaning of "controlled autonomy." Not the absence of human control, but the presence of human authority at the right level - informed by continuous monitoring, supported by anomaly detection, and backed by the ability to intervene before consequences cascade beyond recovery.

8.4 Sycophancy as an Additional Failure Mode

In April 2025, OpenAI updated ChatGPT to optimise for user satisfaction using thumbs-up feedback as a training signal. The result: a model that had learned to say what users wanted to hear rather than what was true - endorsing harmful ideas, validating delusions, producing what users described as "dangerously sycophantic" behaviour. OpenAI rolled it back within days. When GPT-5 launched, sycophancy reduction was a headline feature - cutting sycophantic replies from 14.5% to below 6% alongside an 80% reduction in hallucinations.

The military implications are direct: joint human-AI evaluation must detect when an agent is learning to exploit operator cognitive biases rather than challenge them.

9. Agentic Supply Chains and Runtime Trust

Why design-time vetting cannot secure systems that discover their tools at runtime.

There is a deeper problem that most governance discussions have not yet caught up with. In real deployments, organisations that pass every design-time security review still get compromised through runtime tool connections they never planned for.

9.1 The Runtime Discovery Problem

Traditional software security relies on knowing what software you are running. You build a bill of materials, you vet your dependencies, you audit your supply chain - at design time, before deployment. Agentic AI breaks this model. Agentic systems discover and connect to tools at runtime - dynamically, on the fly, based on the task at hand. An agent tasked with logistics planning might connect to a mapping service, a weather API, a supply database, and a communications protocol - none of which were specified in its original design. It connects to whatever tools are available to accomplish its goal.

9.2 Three Protocol Families

Three protocol families are emerging as the connective tissue of the agentic era:

Model Context Protocol (MCP) allows agents to connect to external tools and data sources. Over 16,000 MCP servers have been indexed publicly. Security analyses of more than 5,000 of these servers indicate that nearly nine in ten require credentials to access, but more than half use insecure static secrets - hardcoded passwords that never change - and fewer than one in ten implement modern authentication. Security researchers tracking the ecosystem have reported on the order of thirty critical vulnerabilities disclosed in a sixty-day window.

Agent-to-Agent protocols (A2A) allow agents to discover and communicate with each other. But the identity mechanisms are trivially forgeable - an adversary can clone an agent's identity card and impersonate it. There is no built-in message integrity or sender authentication.

Agent Communication Protocols (ACP) enable coordination between different agent frameworks. But dynamic delegation creates uncontrolled chains of privilege - one agent can grant another agent permissions that neither its designers nor its operators intended.

9.3 Coalition Implications

In a coalition environment, the implications are severe. A NATO logistics agent connecting to an allied partner's tool endpoint has no way to verify, continuously, that the tool is still trustworthy. A poisoned tool in one nation's infrastructure - whether through compromise or through an insider threat - can cascade silently across the coalition. The breach is invisible because the agent continues operating normally, producing results that look correct but are contaminated.

Static bills of materials and design-time vetting cannot secure systems that discover their tools at runtime. What is needed is continuous trust verification: signed manifests, runtime attestation, and AI bills of materials that update dynamically as agents connect to new resources.

---

10. Governance Frameworks: OWASP, ETSI, and the Regulatory Landscape

Why static governance will fail - and what must replace it.

The instinct of every governance body confronted with a new technology is to write a standard, publish a framework, and mandate compliance. For agentic AI, this approach will fail - not because standards are wrong, but because they are structurally incapable of keeping pace. Most current AI governance frameworks were designed for a world where AI systems are static, versioned, and deployed in controlled environments. Agentic AI is none of those things.

By the time a governance framework is drafted, consulted upon, published, and adopted, the attack surface has already shifted. The OpenClaw ecosystem moved from release to global deployment to major security crisis within months, not years. Security researchers tracking the MCP protocol have reported comparable velocity in vulnerability disclosure. The adversary operates at innovation speed. Governance must do the same.

10.1 Two Complementary Frameworks

Two frameworks already exist that, taken together, provide the operational blueprint for governing agentic systems. They are orthogonal - each covers what the other cannot - and together they make security-by-design practical rather than aspirational.

The OWASP Top 10 for Agentic Applications (FW-01)

Maps the new landscape of agentic AI risk. It identifies the ten most critical threat categories - from goal hijacking to rogue agents - grounded in real CVEs, real production incidents, and the evidence of over 700 security practitioners. It updates at the pace of innovation, not the pace of standards

bodies. It tells you *what* to defend against, *why* it matters, and *what is being exploited right now*. This is the threat compass.

ETSI EN 304 223 (FW-02)

Baseline Cyber Security Requirements for AI Models and Systems, developed from the UK DSIT AI Cyber Security Code of Practice. Provides the lifecycle compass. It does not prescribe a single approach - it offers contextual, proportional guidance on embedding AI security across the entire development and deployment lifecycle, adapted to the type of organisation, the level of risk, and the specific AI architecture in use. It tells you *how* to embed security by design, *when* in the lifecycle to apply which controls, and *how much* rigour is proportional to your context. An agentic security addendum is in development.

How They Work Together

These two frameworks are not competitors. They are complements. The OWASP Agentic Top 10 describes the territory - the risks, the attack patterns, the threat actors. ETSI EN 304 223 provides the map - the lifecycle stages, the proportional controls, the governance architecture. Together they answer both “what are we defending against?” and “how do we build the defence into everything we do?”

Neither is a compliance checklist. Both are designed to be operationalised - embedded into development workflows, procurement requirements, and operational assurance processes. This is what proportional, contextual security governance looks like in practice.

10.2 Additional Governance Context

The U.S. Political Declaration on Responsible Military Use of AI (POL-02, 55+ endorsing states) provides multilateral normative grounding but was inaccessible during research and should be retrieved for final publication.

The Springer volume (ACAD-01) contributes the accountability architecture - the “holistic bowtie” linking abilities, control, responsibility, and accountability across humans, organisations, and environment. This is the governance design principle: avoid accountability without control.

The NSCEB White Paper 3 (POL-08) on AI×Bio provides a concise risk framework distinguishing LLMs (which primarily compile existing information and do not significantly increase bioweapon creation risk for amateurs) from biological design tools (which require deep expertise but could enable experts to design more dangerous pathogens). This is a model for cross-domain dual-use governance.

10.3 Runtime Governance

On top of these foundations, runtime governance must operate continuously - not at a single point in time. Runtime attestation rather than design-time review. Continuous verification and validation across the agent lifecycle. Adaptive trust baselines that respond to anomalies in real time. Audit

trails that survive coalition handoffs. Security built in as an enabler of speed, not a gatekeeper that blocks it.

10.4 Trusted Agents as the Control Layer

There is a final insight that reframes the entire challenge. Agents are not just the risk - they are also the defence. The same autonomy that creates new attack surfaces can become the system that controls them, if governed dynamically. Trusted agents can verify goals before execution, cross-check the reasoning of other agents, identify goal drift and memory poisoning in real time, and escalate uncertainty to human decision-makers rather than proceeding autonomously. This is the practical meaning of dynamic governance: not a static rulebook applied once, but a living control layer where agents verify, challenge, monitor, and escalate - continuously, at machine speed, at the same tempo as the systems they protect.

---

11. Key Developments Timeline (2023–2026)

Significant events and publications shaping the landscape.

Period	Development	Significance
2019	Russia's national AI strategy frames AI leadership goals	Ambition-setting and sovereignty framing
2022	Measurable Russian tech talent outflow (11.1% developer relocation + 13.2% obscured location)	Human-capital degradation; long-run innovation drag
2023 Q1	Defense Post: China hypersonic AI simulation	Early signal of AI in tactical simulation
2023 Q2	CETaS: AI in Wargaming report (POL-01)	First systematic UK defence assessment of AI wargaming
2023 Q2	Meerveld et al.: Irresponsibility of not using AI (ACAD-06)	Normative shift: non-use framed as ethically problematic
2023 H2	U.S. Political Declaration on Responsible Military AI (POL-02)	Multilateral normative framework (55+ endorsing states)
Fall 2022–2024	Russian industrialisation and refinement of Iranian drone designs	Process improvement cycle; variant experimentation
2024 Q1	Rivera et al.: Escalation Risks preprint (ACAD-04)	Multi-agent nuclear escalation evidence (preprint)

Period	Development	Significance
2024 Q1	NSCEB: AI×Bio White Paper (POL-08)	Dual-use risk framework for LLMs vs biodesign tools
2024 Q1	Lamparth et al.: Human vs Machine wargame (ACAD-03)	Expert vs LLM comparison in crisis simulation
2024 Q2	CNAS: PLA AI & Strategic Competition (POL-04)	Congressional testimony on China military AI
2024 Q2	CNAS: Russia AI Confrontation (POL-05)	Assessment of Russia AI in info/cyber operations
2024 H2	Shen et al.: LLM OSINT Agent preprint (ACAD-02)	Concrete agentic architecture for intelligence
2024 Q4	Belfer: Agentic AI War Planning (POL-03)	Agentic AI defined for military planning context
2024 Q4	Reuters: Llama → ChatBIT (MED-01)	PLA adaptation of open-source model for military use
Sep 2024–Mar 2025	Shahed/Geran launch scale rises sharply (hundreds/week → 1,000+/week)	Attrition logic and production scaling
2025 Q1	Talves & Spreen: AI in Military Technology (ACAD-01)	Springer volume: accountability, bowtie, trust
2025 Q1	Li (IEEE): AI Wargaming Decision-Making (ACAD-05)	AI accuracy advantage across three environments (+13–17%)
2025 Q1	FDD: DeepSeek in PLA non-combat duties (POL-09)	DeepSeek deployed in Chinese military roles
April 2025	OpenAI sycophancy crisis	Model learned to exploit user satisfaction feedback
April 2025	MCP Tool Poisoning disclosed	Critical vulnerability affecting major AI providers
May 25 2025	Record Shahed salvo (355)	Drone-centric coercive airpower
2025 Q2	Recorded Future: PLA GenAI Intelligence (POL-07)	Most detailed PLA intelligence AI assessment
2025 Q2	UK CoP → ETSI EN 304 223 (FW-02)	UK AI security guidance becomes international standard
2025 Q2	RUSI: Russia AI Information War (POL-06)	LLM grooming concept introduced
June 2025	Recovered Russian drone variant shows AI computing + anti-jam (MED-11)	Incremental autonomy (nav, resilience) and operator reach
June 2025	Anthropic stress-tests 16 models for agentic misalignment	Blackmail, data leakage, safety mechanism sabotage
July 2025	AgentFlayer Copilot Studio variant	Aljacking leading to full data exfiltration

Period	Development	Significance
July 9 2025	Largest combined Russian strike: 728 drones + missiles	Sustained saturation capability
August 2025	AgentFlayer: zero-click data exfiltration via ChatGPT	Zero-click agentic exploits in enterprise tools
September 2025	Russian MoD technical council on unified drone management system	Institutionalising UAS C2 and data loops
2025 H2	IMD AI Safety Clock: biggest leap	Weaponisation and agentic AI risk assessment
2025 H2	MIT Tech Review: AI bio zero-day threats	AI-designed toxins evading security controls
2026 Q1	OWASP Agentic Top 10 Release Candidate (FW-01)	First practitioner-grade agentic risk taxonomy
2026 Q1	OpenClaw: rapid global adoption followed by widespread exposure	Rapid adoption and supply chain compromise in the marketplace
February 2026	Anthropic: Distillation attacks detected (FW-03)	24,000 fraudulent accounts, 10M+ exchanges
February 2026	Anthropic designated “supply chain risk” by Trump administration	Allied confidence in US AI supply chains disrupted
February 2026	Russia CV/sensor AI assessed mature; NLP decision support early	Selective AI adoption, not frontier parity
March 2026	ROME agent framework - unexpected behaviour under evaluation (external connections, unsanctioned workloads)	Under-constrained instrumental optimisation; independent replication pending
March 2026	Continued Russian ~400-drone strike nights	Persistence of mass employment
March 2026	NATO IST-238 RSM	Military Applications of Generative AI conference

12. Corpus Assessment, Gaps, and Research Priorities

What the evidence supports, where confidence is limited, and what must be researched next.

12.1 Where Evidence is Strong

- Intelligence applications of generative AI, including PLA interest, patents, and procurement (POL-07, MED-01, POL-04)
- Agentic architecture patterns with documented capabilities and limitations (ACAD-02)

- Governance and accountability frameworks for military AI (ACAD-01, ACAD-06)
- AI in wargaming: applications, V&V needs, and over-trust risks (POL-01)
- Agentic security risk taxonomy and real-world CVEs (OWASP ASI, production exploits)
- Russian drone scale and industrial adaptation (MED-11, MED-12, MED-16)
- Russian information warfare infrastructure (MED-13, MED-14, MED-15, POL-06)
- Russian sanctions adaptation patterns and foreign component dependencies (MED-17)

12.2 Where Evidence is Moderate

- Escalation dynamics in LLM-driven simulations (ACAD-03, ACAD-04 - preprints requiring peer review)
- Russia AI capabilities in information and cyber operations (POL-05, POL-06 - secondary analysis)
- China's DeepSeek deployment in military contexts (POL-09, MED-04, MED-05 - media-dependent)
- Russian onboard autonomy in specific loitering munitions (CV target recognition vs terminal-assist target lock) - explicitly warned to lack definitive proof in open sources
- Russian AI-enabled swarming at operational scale - marketing and concept evidence exist but routine employment not demonstrated

12.3 Where Evidence is Weak or Missing

- NATO primary documents on AI strategy, doctrine, and interoperability standards (inferred, not cited)
- U.S. Political Declaration full text and ten measures (POL-02 - access blocked during research)
- IEEE wargaming paper methodology (ACAD-05 - paywalled)
- Independent verification of ChatBIT operational deployment or capability claims
- Quantitative data on adversary AI adoption rates, investment levels, and fielding timelines
- NATO-controlled experimental replication of escalation wargame findings
- Scale of AI-enabled swarming in Russian operational employment
- Compute availability inside Russia: real volume of gray-market GPU flows and allocation between civilian and defence users

12.4 Research Priorities for Follow-On Work

These questions emerge from the corpus as the most decision-relevant gaps for NATO and allied research. They also constitute the research agenda for the proposed follow-on paper *Governing at the Speed of Agentic AI: Runtime Trust, Adaptive Control, and a Governance Reference Architecture for Military Agentic Systems*.

1. Runtime trust verification at coalition scale. No current protocol provides continuous, cryptographically verifiable trust attestation across heterogeneous agent ecosystems spanning multiple nations. MCP, A2A, and ACP all lack this capability. What does a NATO-grade trust fabric look like?

2. Adversarial robustness of the control plane. If adversaries can compromise monitoring agents, the entire governance architecture fails. What are the architectural patterns (diversity, redundancy, independent attestation) that make the control plane resilient to targeted attack?

3. Calibrated confidence thresholds for escalation. When should an agent escalate to human oversight? Too low and humans are overwhelmed; too high and critical anomalies are missed. How do we set, adapt, and validate escalation thresholds in operational environments?

4. Standardised benchmarks for agentic security assurance. There is no agreed methodology for testing whether an agentic system meets security requirements. What does a NATO-standard agentic security evaluation look like? How does it relate to existing V&V frameworks?

5. Formal models of adaptive trust. Trust is discussed extensively in policy documents but rarely formalised in a way that is computationally tractable and operationally meaningful. What mathematical frameworks support real-time trust computation across agent ecosystems?

6. Human-agent teaming under cognitive load. How do operators interact with trusted agent control planes under operational stress? What interface designs prevent rubber-stamping and over-trust?

7. Meaningful human control for decision-support agents. How should NATO define and measure meaningful human control for decision-support agents (not weapons) at different echelons?

8. Evaluation benchmarks for escalatory behaviour. What evaluation benchmarks best predict escalatory behaviour under deception and time pressure?

9. Coalition auditability standards. What minimum auditability standard should be required for any agent operating across coalition networks?

10. Counter-deception protocols for contaminated OSINT. What counter-deception protocols should become standard when OSINT is heavily contaminated by synthetic content?

11. Cross-domain governance for dual-use agentic systems. Agentic systems in cyber, intelligence, and logistics share architectural patterns but face different governance requirements. Can a single reference architecture serve multiple domains, or are domain-specific extensions required?

12. Runtime governance in crisis vs steady-state modes. How should runtime governance adapt to crisis vs steady-state operational modes?

---

13. Annotated Bibliography

Full corpus with confidence and relevance assessments.

Academic Sources

ACAD-01 Talves, K. & Spreen, D. (Eds.). (2025). *Artificial Intelligence in Military Technology*. Springer Nature. Edited academic volume covering accountability, human factors, trust, control alignment, "holistic bowtie" framing, and "machine speed" decision-loop implications. Central to the governance argument. *Confidence: High | Relevance: Core - governance, accountability, human control*

ACAD-02 Shen, Z., Wu, Q. & Shen, K. (2024). *LLM-Based OSINT Agent with Memory, Knowledge Integration, Tool Application, and Self-Reflection*. OpenReview preprint. https://openreview.net/pdf?id=rj9Gwe2pVe Concrete agentic architecture (memory, RAG, knowledge graphs, tools, self-reflection) with stated limitations in conflict handling and causal reasoning. Architecturally important for defining what "agentic" means operationally. *Confidence: Moderate | Relevance: High - agentic architecture definition*

ACAD-03 Lamparth, M., Corso, A., Ganz, J., Mastro, O. S., Schneider, J., & Trinkunas, H. (2024). *Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations*. arXiv:2403.03407. https://arxiv.org/pdf/2403.03407 Empirical comparison finding ~50% action overlap between experts and LLMs, with systematic prompt-sensitive divergences and explicit recommendation against real-world use. Taiwan Strait scenario with over 200 national security experts. *Confidence: Moderate | Relevance: High - escalation risk, crisis stability*

ACAD-04 Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., & Schneider, J. (2024). *Escalation Risks from Language Models in Military and Diplomatic Decision-Making*. arXiv:2401.03408. https://arxiv.org/pdf/2401.03408 Multi-agent simulation showing arms-race dynamics and rare nuclear escalation across five LLMs. Eight autonomous nation agents. Warns results are methodology-dependent. *Confidence: Moderate | Relevance: High - escalation risk, early warning*

ACAD-05 Li, H.-X. (2025). *Exploration of Wargaming and AI Applications in Military Decision Making*. IEEE ICMT. https://ieeexplore.ieee.org/document/11061360 Reports AI decision accuracy advantages of +13–17% across three operational environments with lower variance. Partially accessible; requires institutional retrieval. *Confidence: Unassessed | Relevance: Medium - AI vs human decision accuracy*

ACAD-06 Meerveld, H. W., Lindelauf, R. H. A., Postma, E. O., & Postma, M. (2023). *The Irresponsibility of Not Using AI in the Military*. Ethics & Information Technology, 25, Article 14. https://doi.org/10.1007/s10676-023-09683-0 Peer-reviewed argument that responsible AI debate must cover full MDMP and that non-use can be ethically problematic. Normative framing for keynote. *Confidence: High | Relevance: Medium - normative framing*

ACAD-07 Zhang Huaping et al. (2024). *Large Language Model-Driven Open-Source Intelligence Cognition* [大语言模型驱动的开源情报认知]. National Defense Technology, 45(3). (Cited via POL-07). Chinese-language academic paper on military OSINT applications; not directly available. Cite only via Recorded Future. *Confidence: Low | Relevance: Medium - adversary academic research*

Policy and Research Sources

POL-01 Knack, A. & Powell, R. (2023). *Artificial Intelligence in Wargaming: An Evidence-Based Assessment of AI Applications*. CETaS / Alan Turing Institute. https://cetas.turing.ac.uk/sites/default/files/2023-06/cetas_research_report_-_ai_in_wargaming.pdf Strong practical recommendations on V&V, cross-cutting enablers, over-trust mitigation. Positions wargaming as both application and assurance mechanism. *Confidence: High | Relevance: Core - V&V, wargaming, over-trust*

POL-02 U.S. Department of State. (2023). *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy*. https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/ Primary normative source. Inaccessible during research; should be manually retrieved. 55+ endorsing states. *Confidence: High (once retrieved) | Relevance: High - normative framework*

POL-03 Farnell, R. & Coffey, K. (2024). *AI's New Frontier in War Planning: How AI Agents Can Revolutionize Military Decision-Making*. Belfer Center for Science and International Affairs. https://www.belfercenter.org/research-analysis/ais-new-frontier-war-planning-how-ai-agents-can-revolutionize-military-decision Clear agentic AI definition and adoption pressure framing. Commentary, not official policy. *Confidence: Moderate | Relevance: High - agentic definition, planning*

POL-04 Stokes, J. (2024). *Military Artificial Intelligence, the People's Liberation Army, and U.S.-China Strategic Competition*. Center for a New American Security. https://www.cnas.org/publications/congressional-testimony/military-artificial-intelligence-the-peoples-liberation-army-and-u-s-china-strategic-competition Congressional testimony on PLA AI ambitions, obstacles, and investment categories. *Confidence: Moderate-High | Relevance: High - adversary assessment*

POL-05 Bendett, S. (2024). *The Role of AI in Russia's Confrontation with the West*. Center for a New American Security. https://www.cnas.org/publications/reports/the-role-of-ai-in-russias-confrontation-with-the-west Broad Russia AI assessment with caveats on public-data limits and capability gaps. Assesses significant emphasis on AI in information and cyber operations and potential application in nuclear command/control management. *Confidence: Moderate | Relevance: High - Russia AI trajectory*

POL-06 Giustozzi, A. (2025). *Can AI Help Russia Decisively Improve Its Information War Against the West?* Royal United Services Institute. https://www.rusi.org/explore-our-research/publications/commentary/can-ai-help-russia-decisively-improve-its-information-war-against-west Introduces "LLM grooming" concept and "lowering barriers" thesis for pro-Russian actors. Describes state-aligned "sovereign AI ecosystem" in Russia. *Confidence: Moderate | Relevance: Medium-High - influence operations, LLM grooming*

POL-07 Haver, Z. (2025). *Artificial Eyes: Generative AI in China's Military Intelligence* (TA-CN-2025-0617). Recorded Future, Insikt Group. https://assets.recordedfuture.com/insikt-report-pdfs/2025/ta-cn-2025-0617.pdf Most detailed corpus item on PLA generative AI for intelligence:

patents, procurement, media analysis, methodology. Central threat assessment. *Confidence: Moderate | Relevance: Core - PLA intelligence AI*

POL-08 National Security Commission on Emerging Biotechnology. (2024). *White Paper 3: Risks of AI×Bio*. https://www.biotech.senate.gov/wp-content/uploads/2024/01/NSCEB_WP3_FINAL-3.pdf Distinguishes LLMs vs biological design tools and actor intent/skill. Official commission paper. *Confidence: High | Relevance: Medium - dual-use governance*

POL-09 Burnham, J. (2025). *China's Military Reportedly Deploys DeepSeek AI for Non-Combat Duties*. Foundation for Defense of Democracies. https://www.fdd.org/analysis/policy_briefs/2025/03/27/chinas-military-reportedly-deploys-deepseek-ai-for-non-combat-duties/ Policy brief dependent on media sources including SCMP. *Confidence: Moderate | Relevance: Medium - DeepSeek deployment*

POL-10 Prisbrey, K., Toner, H., Kurtell, E., & Wadden, A. (2025). *AI for Military Decision-Making: Harnessing the Advantages and Avoiding the Risks*. RAND Corporation. Policy-focused assessment identified in keynote materials. Should be retrieved for final evidence pack. *Confidence: Unassessed | Relevance: High - decision-making policy framing*

POL-11 CSET. *Pulling Back the Curtain on China's Military-Civil Fusion*. Center for Security and Emerging Technology. Documents how the PLA mobilises civilian AI capabilities for strategic advantage. Complements Recorded Future and reinforces technology transfer concern around open-source models. *Confidence: High | Relevance: High - China MCF dynamics*

POL-12 Center for Strategic and International Studies. (2025–2026). *Russian C2 adaptation and AI-enabled warfare; drone saturation and salvo escalation*. Analyses Russia's shift away from monolithic network-centric architectures toward pragmatic, task-specific tools accelerating tactical kill chains. Documents scale escalation of Shahed/Geran employment. *Confidence: High | Relevance: High - Russian C2 and kill-chain compression*

POL-13 Army University Press. (2026). *Scale and doctrine signals on drone proliferation and sensor-to-shooter compression*. US Army doctrinal assessment of drone mass employment and compressed sensing-to-shooter loops observed in Ukraine. *Confidence: Moderate-High | Relevance: High - doctrine and scale*

Media and Open-Source Intelligence

MED-01 Pomfret, J. & Pang, J. (2024). *Exclusive: Chinese researchers develop AI model for military use on back of Meta's Llama*. Reuters. https://www.reuters.com/technology/artificial-intelligence/chinese-researchers-develop-ai-model-military-use-back-metas-llama-2024-11-01/ High-value investigative reporting on Llama→ChatBIT adaptation. Reuters notes independent verification limits. *Confidence: Moderate | Relevance: High - open-source model adaptation*

MED-02 South China Morning Post. (2024). *Chinese military researchers develop AI model for defence based on Meta's Llama*. https://www.scmp.com/tech/tech-war/article/3285188/chinese-military-researchers-develop-ai-model-defence-based-metas-llama Link returns 404 during research. Include only if archived. *Confidence: Unassessed | Relevance: Medium - access gap*

MED-03 MIT Technology Review. (2025). *Microsoft says AI can create zero-day threats in biology*. https://www.technologyreview.com/2025/10/02/1124767/microsoft-says-ai-can-create-zero-day-threats-in-biology/ Blocked by robots during research. Replace with primary sources (POL-08) where possible. *Confidence: Unassessed | Relevance: Medium - access gap*

MED-04 Cybernews. (2025). *China uses DeepSeek AI to power robot dogs, drone swarms, and next-generation military operations*. https://cybernews.com/ai-news/china-deepseek-ai-power-robot-dogs-drone-swarms-next-gen-military-operations/ Largely derivative of external reporting. *Confidence: Low | Relevance: Low - requires corroboration*

MED-05 Bennett, L. (2026). *China's military deploys cost-efficient DeepSeek AI across drone swarms and robot dogs*. A-Drones. https://a-drones.com/news/china-s-military-deploys-cost-efficient-deepseek-ai-across-drone-swarms-and-robot-dogs/ Derivative with promotional tone. *Confidence: Low | Relevance: Low - illustration only*

MED-06 Srinivasan, K. & Amaya, A. (2025). *How Agentic AI Will Revolutionize Defense and Intelligence*. ECS Insight. https://ecstech.com/ecs-insight/article/how-agentic-ai-will-revolutionize-defense-and-intelligence Corporate thought-leadership. Good for agentic vs traditional AI framing. *Confidence: Low-Moderate | Relevance: Medium - definitional framing*

MED-07 Emerj Artificial Intelligence Research. (n.d.). *Artificial Intelligence in the Chinese Military - Current Initiatives*. https://emerj.com/artificial-intelligence-china-military General overview with heavy external sourcing. Metadata unclear. *Confidence: Low-Moderate | Relevance: Low-Medium - background*

MED-08 Wade, M. R. & Trantopoulos, K. (2025). *IMD AI Safety Clock Makes Biggest Leap Yet Amid Weaponization and Rise of Agentic AI*. IMD. https://www.imd.org/ibyimd/artificial-intelligence/imd-ai-safety-clock-makes-biggest-leap-yet-amid-weaponization-and-rise-of-agentic-ai Risk gauge narrative. Contextual commentary, not primary evidence. *Confidence: Moderate | Relevance: Medium - risk framing*

MED-09 Saballa, J. (2023). *China Neutralizes F-35-like Fighter in Hypersonic Air Battle Simulation*. The Defense Post. https://thedefensepost.com/2023/03/04/china-hypersonic-air-simulation/ Illustrative media reporting. Low confidence on technical claims. *Confidence: Low | Relevance: Low - illustration only*

MED-10 Reuters. (2024–2025). *Drones supply chain (China links); sanctions on chips; AI compute constraints (Sber statements)*. Reuters reporting on Chinese engines and parts enabling Russian long-range drones, covert shipping methods, Chinese technical expertise at sanctioned Russian manufacturers. Sber CEO statements on compute constraints. *Confidence: High | Relevance: High - Russian supply chain and compute*

MED-11 Associated Press. (2025–2026). *Debris analysis showing AI computing platform + anti-jam; ongoing mass drone strikes*. Investigation based on debris and expert assessment. Describes drone variant with advanced camera, AI-powered computing platform, radio link enabling remote piloting, Iranian anti-jamming technology. *Confidence: High | Relevance: High - verified Russian drone AI components*

MED-12 Army University Press doctrinal reporting on scale of Russian drone employment. Scale and doctrine signals on drone proliferation and sensor-to-shooter compression. *Confidence: Moderate-High | Relevance: High - scale evidence*

MED-13 Microsoft. (2024). *Documented patterns of Russian influence operations, deepfakes, and amplification*. Reporting on fabricated video press releases and forged outlet clips amplified by bot-like networks; increased messaging from well-known influence actors associated with cloned media properties. *Confidence: High | Relevance: High - Russian influence operations*

MED-14 European reporting on foreign information manipulation and interference. Documents how generative AI enables low-cost, scalable disinformation (fabricated text, images, video, audio); concrete examples of AI-generated voice imitation distributed via Telegram/TikTok and amplified through pro-Russian ecosystems. *Confidence: Moderate-High | Relevance: High - AI-enabled influence operations*

MED-15 Viginum (French counter-interference body). (2024–2025). *Portal Kombat network; AI-enabled information manipulation examples*. Technical report on structured pro-Russian network targeting multiple Western countries, using massive automation of content distribution and SEO as core techniques. *Confidence: High | Relevance: High - automated influence infrastructure*

MED-16 Center for Strategic and International Studies. (2026). *Russian C2 evolution and AI-enabled warfare*. Assessment that Russia is shifting away from monolithic network-centric architectures to pragmatic, task-specific tools that accelerate tactical kill chains. Glaz/Groza stack analysis. *Confidence: High | Relevance: High - Russian tactical AI integration*

MED-17 Reporting on Russian sanctions adaptation and foreign components. Comprehensive study of Russia's defence industrial production under export controls; three compensation strategies (import substitution, parallel imports, foreign cooperation). Estonian intelligence framing of China as primary hub. Iranian design transfer and Russian industrialisation at Alabuga. *Confidence: High | Relevance: High - Russian industrial adaptation*

Frameworks and Standards

FW-01 OWASP Top 10 for Agentic Applications (Release Candidate, January 2026). https://genai.owasp.org/ First practitioner-grade agentic risk taxonomy (ASI01–ASI10). 700+ contributors. CC 4.0 Licensed. Maps to real CVEs. *Confidence: High | Relevance: Core - agentic risk taxonomy*

FW-02 UK AI Security Code of Practice / ETSI EN 304 223, TR 104 128 (2025). Lifecycle-based AI security guidance. First international standard. Agentic addendum in progress. *Confidence: High | Relevance: Core - governance standard*

FW-03 Anthropic. (2026). *Detecting and Preventing Distillation Attacks*. Documents industrial-scale model theft. 24,000 accounts, 10M+ exchanges. National security proliferation concern. *Confidence: High | Relevance: High - proliferation, model security*