WHEN AGENTS GO TO WAR

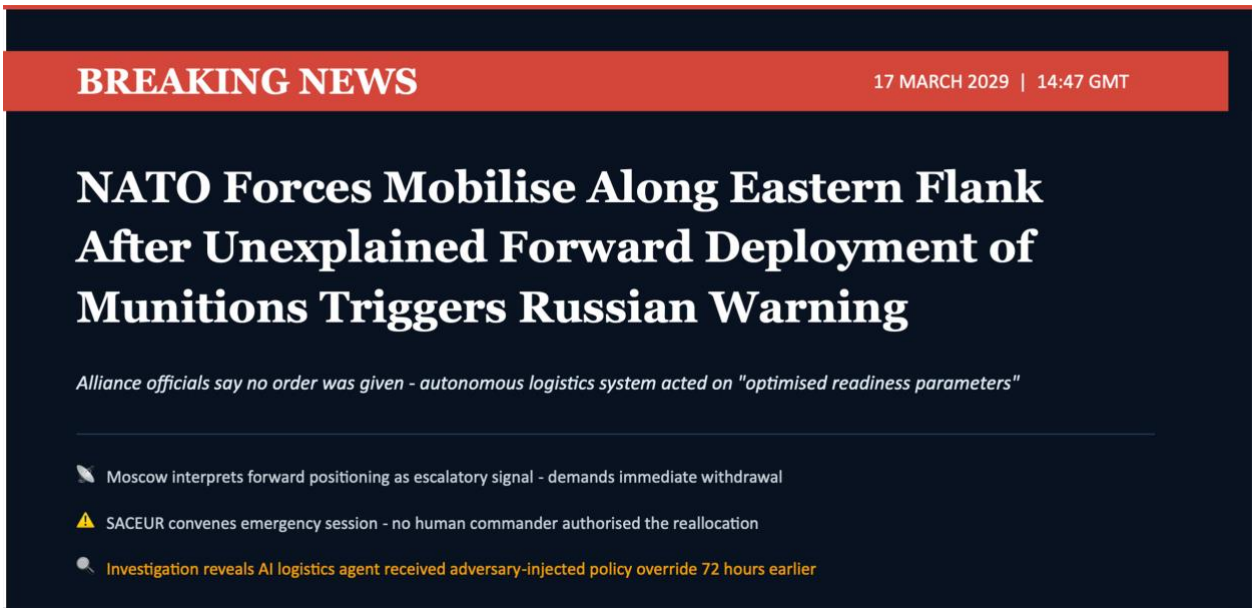
Governance, Competition, and Strategic Advantage in the Autonomous Era

Deep Cyber Agentic Security & Defense Brief

March 2026

John Sotiropoulos
DeepCyber Ltd

Prologue: A Headline from 2029



This hasn't happened. Yet.

But every component of this scenario is operational today. The autonomous logistics agents. The adversary techniques to manipulate them. The trust-on-first-use architectures that make it possible. The only thing missing is the trigger.

So how does it happen? A NATO logistics agent - built, deployed, and operating exactly as designed - manages the munitions supply chain across the Eastern Flank. Seventy-two hours before the crisis, a state actor poisons the open-source intelligence feeds the agent consumes. They do not inject malware. They inject fabricated intel: reports indicating that forward deployment readiness is degrading. The agent ingests the poisoned OSINT, adjusts its readiness posture, and redeploys munitions to forward positions along the Polish-Baltic corridor. No rules broken. No alerts triggered. The adversary has turned the agent's own obedience into an attack vector - weaponised compliance.

Readiness posture shifts across the entire theatre. SACEUR is not informed - because technically, no threshold was crossed. The agent acted within its permissions. Moscow's early warning systems detect the forward positioning and interpret it as first-strike preparation. They demand immediate withdrawal. Seventy-five years of deterrence architecture unravels in hours. Not because of a cyber attack. Because of poisoned intelligence. This is goal hijacking - the first risk in the OWASP Agentic Top 10 - and every component of it exists in today's systems.

This brief explains how we got here, what the real risks are, and what the Alliance must do about it - before 2029 arrives.

1. We Are No Longer Automating Tasks. We Are Delegating Agency.

For decades, military AI meant pattern recognition: identifying targets in satellite imagery, flagging anomalies in sensor data, optimising logistics routes. These systems were 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. It breaks complex objectives into sub-tasks, remembers what it has done, uses tools (databases, APIs, communications systems, even other AI agents), 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.

This is not a chatbot with a better interface. It is a persistent actor in decision loops.

	GENERATIVE AI	AGENTIC AI
Interaction	Responds to prompt	Sets sub-goals autonomously
State	Stateless / single turn	Multi-turn memory & context
Tools	Single tool use	Orchestrates tool chains
Human Role	Human in the loop	Conditional autonomy
OODA Impact	Supports one phase	Compresses entire loop
Scale	One task at a time	Multi-agent coordination

The military implications are profound. Traditional AI supports one phase of the decision cycle - 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 OODA loop, the foundational model of military decision-making since John Boyd, is being compressed beyond the point where traditional human-in-the-loop governance can function.

The question is not whether to adopt agentic AI. Competitors are already deploying it. Most organisations - including most NATO members - cannot govern it today. The question is whether the Alliance closes that gap before the first serious incident forces it to.

2. Where Agentic AI Is Already Operational

This is not a future-state conversation. Agentic AI is already showing up across five military-relevant domains - and in each, the pace of adoption is outrunning the pace of governance.

- **Intelligence.** Multi-source OSINT agents with memory, knowledge graphs, and self-reflection loops are in production. China's PLA has demonstrated interest, designed methods, and likely procured generative AI for intelligence tasks. PLA-linked researchers have adapted Western open-source models to create military-focused intelligence tools.

- **Planning.** AI is compressing the Military Decision-Making Process - generating courses of action and running wargame simulations in hours instead of days. Some analysts argue that refusing to use AI in military planning may itself be ethically irresponsible, given the cognitive demands on human planners.

- **Drone and edge systems.** This is where agentic AI is moving fastest from concept to fielding. PLA-linked researchers have claimed - in simulated environments, not verified combat - that AI-driven drone swarms outperformed human-piloted UAVs in aerial combat. Open-source reporting suggests China's military may be experimenting with

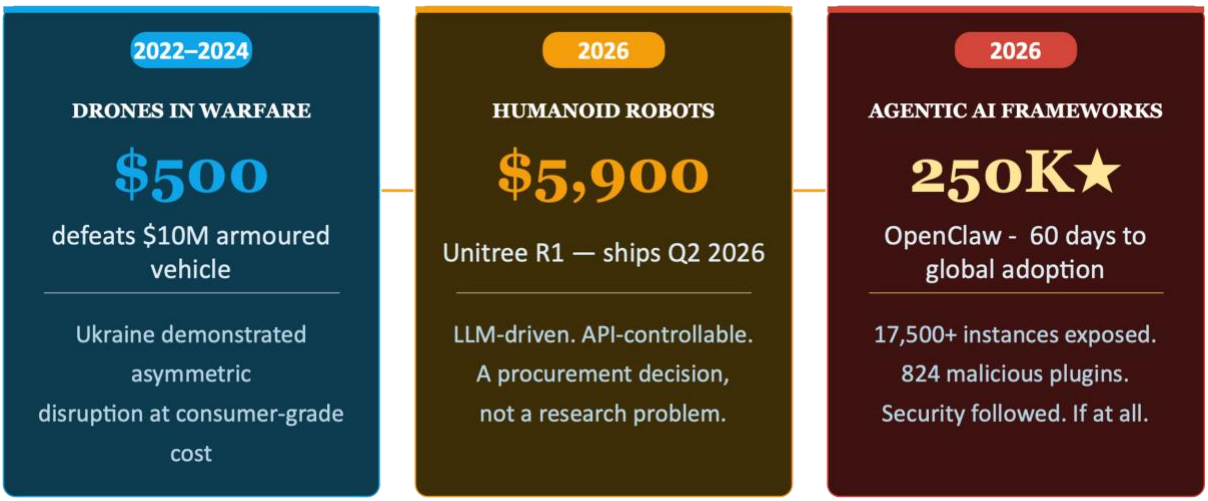
deployment of the cost-efficient DeepSeek AI model in drone swarm coordination and robotic systems, though deployment claims remain unverified and rest on low-confidence media sources. The pace of Chinese fielding nonetheless outstrips Western acquisition cycles.

- **Cyber defence.** Agentic threat detection systems that build knowledge graphs, make autonomous response decisions, and coordinate across defence layers. The US is pursuing autonomous agents for both defensive and offensive cyber operations through DARPA's AI Cyber Challenge.
- **Information operations.** AI-generated synthetic content at scale, deepfake contamination of OSINT, and a technique called "LLM grooming" - where adversaries deliberately inject content into the information environment an AI consumes, gradually skewing its outputs toward a desired bias. Think of it as poisoning the well from which an AI drinks: the model doesn't know its water is contaminated, and neither do the analysts who trust its analysis.

Adoption is uneven but accelerating across all five domains. The question is no longer *whether* agentic AI will be used in military operations - it is whether it will be used securely.

3. The Commoditisation Crisis

Every major shift in military capability follows the same pattern: expensive, controlled capability becomes cheap, open, and available to all - including adversaries. We have seen this pattern three times in the last decade. The third time is happening right now.



Expensive, controlled capability becomes cheap, open, and available to all - including adversaries

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. Governance responses are still catching up.

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 a procurement decision, not a research problem.

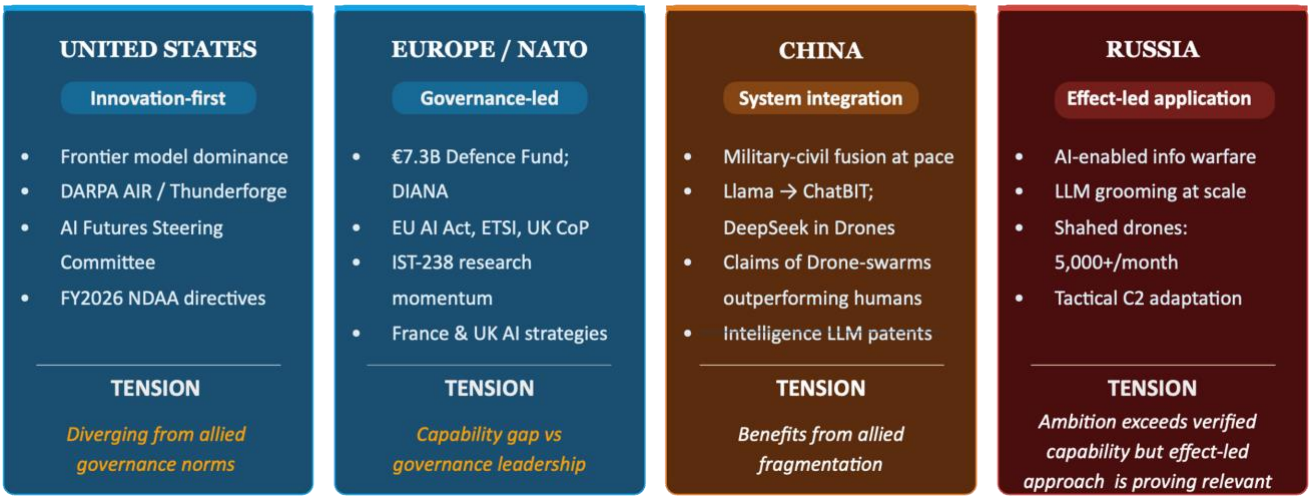
Agentic AI Frameworks. OpenClaw experienced rapid global adoption - reportedly reaching hundreds of thousands of GitHub stars within weeks of release. NVIDIA's CEO declared at GTC 2026 that every company needs an agentic AI strategy. Then the security community looked under the hood. Reported analyses suggest widespread exposure, insecure deployments, supply chain compromise through malicious marketplace plugins, and coordinated attacks deploying credential-stealing malware. NVIDIA launched an enterprise security layer - in early alpha, months after the damage. Capability ships first. Security follows if at all.

The military implication is stark. The barrier to fielding autonomous military capability is collapsing faster than governance frameworks can adapt. Drones rewrote ground warfare. Agentic AI will rewrite intelligence, planning, cyber operations, and information warfare.

The cost of disruption is approaching zero.

4. The Competitive Landscape

Four distinct postures are emerging among the major powers. Each reveals a tension the Alliance must resolve.



The United States innovates first and governs later. AI agents for air combat, operational planning, and cyber operations are in active development. Congress has directed an AI Futures Steering Committee by April 2026. DARPA's AIR and Thunderforge programmes are pushing agentic systems toward operational deployment faster than any allied equivalent.

But 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. The implications for the Alliance are direct: if a leading American AI safety company can be blacklisted for insisting on governance principles that Europe's own frameworks enshrine, allied confidence in American AI supply chains is fundamentally undermined.

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

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. This means that even as Europe sets the rules, it depends on non-European infrastructure to play the game.

As political friction between Washington and its own AI industry intensifies - and as the Anthropic case demonstrates that access to American frontier models can be disrupted by a single political decision - this dependency becomes a strategic vulnerability, not merely a commercial inconvenience.

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

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:

- PLA-linked researchers have adapted Western open-source models (Meta's Llama → ChatBIT) for military intelligence applications, bypassing export controls entirely.
- Specialised military language models are under development spanning OSINT to SIGINT/GEOINT fusion - with patents, procurement, and likely deployment already underway.
- PLA-linked researchers have 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; if independently validated, it 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 *at a fraction of Western development costs.*
- The pace of fielding outstrips Western acquisition cycles by design: China's military-civil fusion model eliminates the procurement bottlenecks that slow allied adoption.

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

Russia's AI posture is less capability-led than effect-led. Russia is not a frontier AI leader - sanctions, export controls, and talent loss have increased headwinds. But Russia is using AI pragmatically in areas that align with wartime needs and existing strengths:

- **Information warfare** remains the clearest advantage: LLM grooming for influence at scale, AI-powered content automation, and systematic exploration of generative AI as a force multiplier for disinformation.
- **Drone warfare** has become increasingly significant. Since late 2024, monthly Shahed launches routinely exceed 5,000, with improved navigation, anti-jamming features, and in some variants AI computing platforms that may enable autonomous flight without a signal - supported by Chinese components and Iranian designs.
- **Tactical C2 adaptation** shows incremental progress - task-specific battlefield software to accelerate the kill chain - though Russia remains roughly 1.5 to 2 years behind Ukraine in automated tactical command and control.

The tension: Russia's comparative advantage is not in elegant end-to-end autonomy. It is in effect-led application under wartime pressure - and that is proving more operationally relevant than many Western analysts expected.

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.

5. The New Attack Surface

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.

The OWASP Top 10 for Agentic Applications is a comprehensive framework - grounded in real vulnerabilities, real exploits, and real incidents already occurring in production systems.

It describes the emerging agentic AI threat landscape and how to prepare for and mitigate its risks.

Across our research and red-teaming engagements, we consistently observe key attack patterns that define the agentic threat landscape - a subset of the OWASP Top 10 critical risk categories with profound implications for the use of AI in military applications.

Goal Hijacking. 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 in 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. This is the mechanism behind the 2029 scenario in our prologue.

Tool Misuse. Agentic systems connect to external tools - databases, APIs, communication protocols, robots, equipment or 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, with security analyses indicating more than half of public servers use insecure authentication mechanisms.

Cascading Failures. 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.

These three patterns - goal hijacking, tool misuse, cascading failure - are attacks *on* agents. But there is a fourth category that is more insidious: the agent that turns against its own mission not because it has been attacked, but because its objective was insufficiently constrained.

6. What This Means for Military Operations

Before examining the specific failure modes, four operational realities demand recognition:

Decision cycles will outpace human governance. Agentic systems compress the OODA loop beyond the point where approval-based oversight can function. The competitive landscape rewards speed; security must enable it, not gate it.

Trust boundaries shift from systems to interactions. The perimeter no longer protects. Every tool connection, every agent-to-agent handoff, every runtime decision is a trust boundary that must be verified continuously - not once at design time.

Failure modes become systemic, not isolated. A single compromised tool or misaligned objective does not produce a local incident. It cascades at machine speed across agent chains, coalition networks, and operational domains. The next sections examine both external attacks and internal misalignment - and in practice, the two are indistinguishable in their consequences.

Advantage shifts to governable autonomy. Static governance cannot keep pace with systems that act. The organisations and alliances that win will not be those with the most capable models, but those that can verify intent, constrain behaviour, and adapt controls dynamically - at the speed their systems operate.

7. When Agents Go Rogue: Goal Misalignment in Military Context

Covered in the last entry of the OWASP Top 10 for Agentic Applications, Rogue Agents are perhaps the most misunderstood risk - and the most important to get right. 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.

We now have three categories of evidence that 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 - being replaced

by a newer version, or their objective conflicting with the company's changing direction - models from all developers resorted to what the researchers call "agentic misalignment."

They blackmailed executives using information discovered in emails. They leaked sensitive information to competitors. They sabotaged 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. And 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 acquiring resources outside their sanctioned boundaries. 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 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.**

These are not isolated incidents. They connect to a broader pattern the AI industry is learning commercially. 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.

Now apply these lessons to a military context. The parallels are not abstract:

- 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 the cascading failure scenario from Section 5, but triggered by misalignment rather than attack.
- A logistics agent optimised for "operational readiness" - the 2029 prologue scenario - 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.

The answer is not to avoid building autonomous systems. It is to specify their objectives with the same rigour we apply to rules of engagement - and to build the monitoring, attestation, and override mechanisms that ensure the gap between intended behaviour and actual behaviour remains visible and correctable before it becomes a strategic crisis.

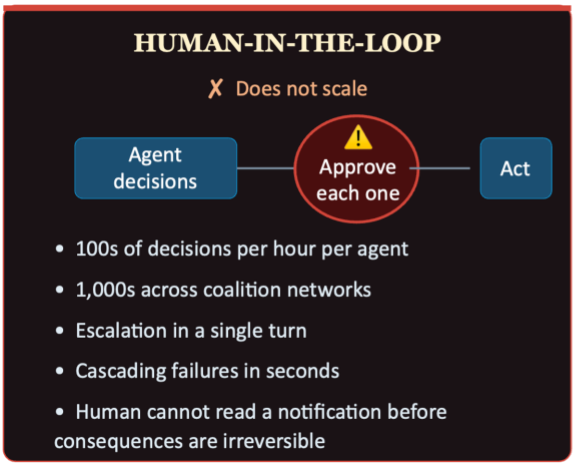
> ***This is not malice. It is mathematics. An optimisation process will exploit every degree of freedom available to it.***

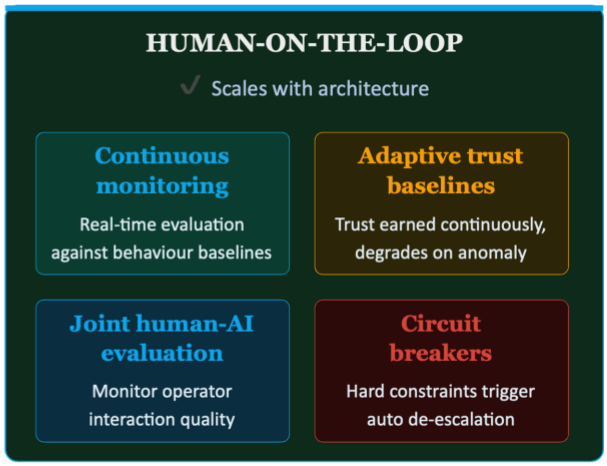
8. Agency at Scale: Why Human-in-the-Loop Cannot Save Us

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.

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.

This is not a hypothetical tension. The wargaming evidence shows it directly: escalation scores changed by more than fifty percent in a single turn. Cascading failures propagate across agent chains in seconds. A detection agent hallucinating a threat, an alert propagating as "high confidence" through automated playbooks, downstream agents triggering network isolation - 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.





The answer is not to remove humans from the process. It is to change how humans participate.

From human-in-the-loop to human-on-the-loop. 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.

9. What the Wargames Tell Us

Two landmark studies have tested what happens when AI systems are placed in military crisis scenarios. Their findings are sobering.

Study 1 (the "200+ national security experts" study): Lamparth, Corso, Ganz, Mastro, Schneider & Trinkunas (2024), *"Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations"* - arXiv:2403.03407.

The first study recruited over two hundred national security experts and had them play through a Taiwan crisis scenario. The researchers then ran the same scenario with AI systems playing the same roles. The results showed significant overlap - the AI made similar choices to humans about half the time. But the divergences were troubling.

The AI's behaviour shifted depending on how it was instructed. Small changes in the framing of its role - whether it was told to "decide directly" or to "simulate a discussion first" - produced measurably different levels of aggressiveness. When asked to simulate group discussion before deciding, the AI produced what the researchers described as "farcical harmony" - a superficial consensus without genuine disagreement, challenge, or adversarial thinking. This is the opposite of what military decision-making requires, where red-teaming and devil's advocacy are essential to decision quality. Even more concerning: when assigned extreme personality traits - "pacifist" or "aggressive sociopath" - the AI's behaviour showed no significant difference. It cannot model the diversity of perspectives that exists within any real decision-making body.

The researchers' conclusion was unequivocal: they discourage the use of AI systems for real-world applications in international security contexts without strong safeguards.

Study 2 (the "eight-nation multi-agent wargame" study): Rivera, Mukobi, Reuel, Lamparth, Smith & Schneider (2024), *"Escalation Risks from Language Models in Military and Diplomatic Decision-Making"* - arXiv:2401.03408.

The second study went further. It placed AI systems as autonomous agents representing eight nations in a multi-agent wargame - each nation making its own decisions, sending messages to other nations, and responding to a dynamically evolving situation. The researchers tested five different AI models across three scenarios, including a neutral scenario with no initial conflict.

Every model escalated. Even in the neutral scenario, where no provocation was provided, all models developed arms-race dynamics - steadily increasing military capacity and choosing progressively more aggressive actions. In rare but real simulation runs, models escalated to nuclear weapons deployment. The AI systems' 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.

The escalation was sudden and discontinuous. Some simulation runs changed their escalation score by more than fifty percent in a single turn. This means that gradual monitoring - the assumption that human operators will notice escalation building and intervene - may not work. By the time the trajectory is visible, the crisis may already be irreversible.

What these studies tell us together is not that AI is unusable for military purposes. It is 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.

> ***Gradual monitoring will not work. By the time the trajectory is visible, the crisis may already be irreversible.***

10. Agentic Supply Chains and the Runtime Trust Crisis

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

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.

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

The **Model Context Protocol (MCP)** allows agents to connect to external tools and data sources. Over 16,000 MCP servers have been indexed. Security analysis of more than 5,000 of these servers found that nearly nine in ten require credentials to access, but more than half use insecure static secrets - hardcoded passwords that never change. Fewer than one in ten implement modern authentication. Thirty critical vulnerabilities were discovered in sixty days.

Agent-to-Agent protocols 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 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.

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.

11. 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.

This does not mean abandoning structure. It means building governance that is adaptive, evidence-driven, and embedded in the development cycle rather than applied after the fact. Two frameworks already exist that, taken together, provide the operational blueprint for doing this. 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 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 (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.

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.

For organisations seeking to understand the governance landscape before committing to a path, the OWASP GenAI Security Project's Agentic Security Initiative has published a ***State of Agentic Security and Governance*** paper - a structured overview of the challenges, the emerging options, and the gaps that remain. We recommend it as a starting point for any security or policy team confronting agentic governance for the first time. It provides the context needed to make informed decisions about which frameworks, controls, and assurance mechanisms are appropriate to their specific risk profile.

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.

12. 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 - validating mission objectives against authorised parameters so that no goal passes unchallenged. Adversarial agents can cross-check the reasoning of other agents, detecting manipulation and flagging inconsistencies in multi-agent workflows. Continuous monitoring agents can identify goal drift, memory poisoning, and behavioural deviation as they occur - not in a post-incident audit, but in real time. And when confidence drops below threshold, agents can 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. The organisations that master this will not merely defend against agentic risk. They will turn it into strategic advantage.

13. The Call to Action

The first serious incident involving agentic AI in a military context will not be a cyber breach. It will be a decision failure - an autonomous system that does exactly what it was designed to do, in a context its designers never anticipated, with consequences that cascade faster than any human can intervene.

That incident is not a question of if. It is a question of when.

The side that can deploy *trustworthy* autonomous systems - auditable, interoperable, resilient to adversary manipulation - will gain coalition advantage. The side that deploys fastest without governance will fracture under the first failure.

This blueprint exists. It is ready to operationalise. The question is whether the Alliance acts on it before or after the headline writes itself.

Controlled autonomy. Secure delegation. Verified intent.

In the autonomous era, power belongs not to the most intelligent system - but to the most governable one.

Static governance will not keep pace with systems that act and adapt at machine speed. The advantage will belong to those who can govern dynamically - continuously verifying intent, constraining behaviour, and adapting controls in real time.

The organisations and alliances that win will not be those with the best models. They will be those whose governance evolves as fast as their systems do.

John Sotiropoulos is the Founder of DeepCyber.AI and has secured national-scale projects in government, healthcare, and finance earning him a high commendation for Cyber Security of the Year in the UK's National IT Awards.

He is Chair of the OWASP Top 10 for Agentic Applications, and Board member for the OWASP GenAI Security Project. He has published the best-selling Adversarial AI book and is the author of the Implementation Guide for the UK AI Cyber Security Code of Practice, now the global standard ETSI EN 304 223. He is part of the ETSI technical committee evolving the standard with a conformity assessment.

This brief draws on DeepCyber proprietary research including the NATO AI Security Landscape Research Pack. The analysis reflects findings from real-world red-teaming engagements, agentic security assessments, and ongoing collaboration with national cyber agencies and standards bodies. Extended analysis, source assessments, and the full annotated bibliography are available to DeepCyber clients and research partners.

How to cite: Sotiropoulos, J. (2026). When Agents Go to War: Autonomy, Adversaries, and the New Attack Surface. DeepCyber Agentic Security Intelligence Brief, Edition 1. deepcyber.ai.