Militaries Are Buying AI Agents Faster Than They Can Assure Them

Test and evaluation rest on assumptions that agentic systems weaken by design. Restoring justified confidence will require new methods.

Ulysse Richard[1], Sarah Cao[1,2], Di Cooke[1,3], Sebastian Kwon[1,4], Adrianna Tan[1,5] · August 2026

[1] Arcadia Impact, [2] Oxford University, [3] King's College London, [4] Atlantic Council, [5] Future Ethics Lab

Summary

Problem: Militaries are procuring AI agents for command and control under public commitments to rigorous testing, but established testing and evaluation methods rest on traditional assumptions that do not hold under agentic conditions.

Why it matters: Testing results may satisfy procedural requirements yet fail to provide fielding authorities with actionable evidence of the system's operational behavior, weakening the basis for justified confidence in fielded performance.

Argument: Agentic properties erode eight assumptions of current test and evaluation practice, including that system behavior is specifiable in advance, that such characterization remains stable across repeated trials and throughout the system's lifecycle, that evidence from components persists through integration, and that operators retain effective supervisory control. This complicates the assurance argument that connects claims to supporting test evidence.

Recommendations: Restoring justified confidence requires aligning assurance claims with the scope of available evidence; developing new testing methodologies that address agentic properties and treat fielding determination as a continuing act; and establishing mechanisms to explicitly manage residual risk that remains unevidenced.

A promise that needs a method

In June 2026, the United States Department of War announced Agent Network, a capability employing AI agents to scan defense intelligence and operational systems and present commanders with options. The announcement commits to rigorous testing, operational evaluation, and oversight throughout development and fielding.

Testing and evaluation (T&E) is the process of gathering evidence about system behavior to determine fitness for purpose. The decision to field a system is typically based on an assurance case, which includes three elements: claims that specify the conditions for acceptability, evidence that bears on those claims, and an argument that connects the evidence to the claims.

In military systems, assurance claims address multiple dimensions, including functional and non-functional performance, robustness to both natural and adversarial conditions, safety, and human-machine teaming. The T&E lifecycle typically follows a pyramid structure: it begins with testing the smallest components, proceeds to integration into larger systems, and culminates in evaluation under operational conditions.

Agentic systems possess inherent properties that resist traditional testing and evaluation. These systems pursue assigned goals without prescribed procedures, operate autonomously in their environment using tools and memory, and adapt dynamically to evolving circumstances.

From a commander's perspective, agentic properties offer opportunities to enhance command and control (C2) functions, the exercise of authority and direction over assigned forces. Effectiveness gains are possible throughout the C2 lifecycle, from improving the completeness, timeliness and relevance of information that feeds into distributed situational awareness to advancing decision quality, timeliness, and synchronization.

But C2 workflows are also highly dynamic, subject to adversarial interference, and dependent on tacit, judgment-based tasks that require close collaboration with human operators. Agentic systems disrupt more fundamental C2 tenets, such as allocation of decision rights (who decides what), patterns of interaction among actors (who communicates with whom and how), and distribution of information (who knows what). Their integration therefore challenges deeper assumptions about the structure and content of an assurance case.

Challenging testing and evaluation assumptions

Traditional T&E practice rests on several clusters of assumptions for generating and interpreting evidence:

- Specifiability: that system inputs, action space and integration surface are reasonably specifiable in advance;
- Stability: that such characterization remains stable across repeated trials and throughout the system's lifecycle;
- Composability: that evidence from components persists through integration;

- Supervisability: that operators retain effective supervisory control.

The unit of behavior shifts from output to trajectory. Established methods evaluate outputs, but for agentic systems, the relevant metric is the trajectory, from the sequence of tools invoked to the data retrieved and written to memory. Focusing solely on the final outcome can obscure failures detectable only by examining the process. The test article is also inherently dynamic. For example, a system that maintains the maritime picture across watch rotations propagates its own inferences together with the original observations, without distinguishing between them. As a result, the system behaves differently from the one originally characterized, current practice does not trigger revalidation.

Certain forms of evidence can only be generated at runtime. The nature of agentic systems and the contested environment of C2 make it infeasible to fully specify a system's inputs, action space, and integration surface in advance. Approaches such as behavioral monitoring, re-baselining, goal-drift detection, and staged fielding transfer part of the evidentiary burden from design time to runtime. As a result, fielding determination becomes a continuous process, with analogs in domains like aviation and nuclear power where re-accreditation is already linked to observed behavior.

Multi-agent systems are not reducible to the sum of their parts. Staged testing presumes that the interface between two components is a specification fixed at design time. In contrast, interfaces between agents are outputs generated at runtime by one agent and interpreted by another; couplings thus emerge only during system operation. This allows behaviors to arise in the assembled system that component-level testing does not reveal, such as individually valid outputs that conflict when combined or errors that gain authority as they propagate through multiple agents.

Oversight is assumed without sufficient evidence. For supervision to be effective, the operator must maintain an accurate operational model, possess the authority to intervene, and have sufficient time to act before outcomes are realized. Factors such as system opacity, adaptive behavior, and agent-to-agent delegation undermine these conditions by expanding the supervisory span and reducing the available intervention window, particularly under high operational tempo. As system reliability increases and operator approval becomes routine, decision-making effectively moves to the system, regardless of formal process. The presence of an operator does not constitute meaningful oversight.

Recommendations

These challenges do not make the systems unusable or testing futile. However, assurance claims must be calibrated to the scope of available evidence. Three near-term approaches merit consideration: narrowing claims, developing new methods, and fostering collaboration across relevant communities of practice.

Narrower claims. Broad claims that a system is safe for command and control are not supportable on the documented record. Narrower claims can be substantiated in principle, each contingent on methods that remain under development. Examples include testing within a bounded mission envelope that defines an agent's trusted role, or applying formal methods to test and enforce constraints prior to high-consequence actions.

New methods. Agentic properties require developing new testing and evaluation paradigms. Open areas for evaluation include characterizing variance over operational trajectory lengths, assessing memory and accumulated state, employing LLM-as-a-judge approaches, and conducting red teaming of multi-agent systems.

Shared practice. The relevant methods and decision-making authority are distributed across different communities. Military T&E and AI evaluation communities often address the same assurance challenges from separate perspectives. While military T&E brings established assurance frameworks, lifecycle processes, and operational testing experience, AI evaluation contributes much of the recent progress in understanding system trajectories, memory, and multi-agent dynamics. Initiatives such as joint workshops, shared terminology and evaluation results designed to be interpretable across both domains would strengthen near-term assurance.

While none of these gaps are insurmountable, addressing them requires time that current procurement schedules often do not allow. Ultimately, whether progress is made on these issues before systems are fielded or addressed afterward is a matter of present decision-making.