When Human Approval Fails: Runtime Authority Governance for Autonomous Deterrence

By
Burak Oktenli

Published: November 5, 2026

A commander approves an autonomous defensive engagement while the target is inside an authorized area. Two seconds later, before the system acts, an allied aircraft enters the area, or the target moves beyond the permitted boundary. The approval is genuine, but does it still authorize the action?

This is a hypothetical sequence, not a documented battlefield failure. It exposes a neglected distinction in military artificial intelligence (AI). Recording a human decision is different from ensuring that the decision remains valid when an autonomous system carries it out.

The timing matters because military AI is moving closer to consequential decisions. The January 2026 Department of War AI strategy calls for AI-enabled battle management and decision support, while the Air Force AI strategy seeks shorter sensor-to-shooter timelines. The June national security AI memorandum emphasizes the chain of command and orders updated autonomy guidance. NATO's 2026 digital strategy likewise calls for integrating sensor-to-effector data flows and building federated platforms for allied collaboration.

None of these strategy documents requires proof that a deployed weapon has been given authority. Instead, they focus on urgency. How can forces accelerate legitimate action without allowing yesterday's permission to survive the facts that justified it?

Human judgment can fail at a seam between the decision and its execution. An operator might authorize a particular defensive action based on a target classification, a geographic limit and evidence that friendly forces are clear. If the conditions change while the action is pending, software that checks only whether an approval flag exists may treat obsolete authorization as current.

A digital signature can authenticate the credential used to approve an instruction. It cannot, by itself, establish that the relevant sensor evidence, rules, or mission limits have not changed since approval. An accurate model prediction does not answer that authorization question either.

The author's own bounded formal analysis, using the Temporal Logic of Actions, illustrates this failure path in notional authority designs. Confirmation remains recorded while a material change invalidates the conditions behind it. Requiring the approval to remain tied to the current context removes that particular modeled violation. This is a result inside a constructed specification, not evidence that actual military control software contains the same weakness.

The public autonomy-in-weapons directive already provides a policy foundation. It requires appropriate human judgment and expects covered systems to remain within authorized time, geographic and operational limits. When that is not possible, the directive calls for termination of

the engagement or additional operator input, as applicable. A force that cannot conduct authorized defense quickly invites doubts about its ability to protect allies. A force that conducts an engagement after its permission has become invalid invites a different doubt. It questions whether commanders can still control escalation. Credible deterrence requires capability and restraint to control escalation, not speed at the expense of either. This is particularly consequential during a conventional confrontation involving nuclear-armed powers. A mistimed or misdirected engagement could be interpreted as deliberate escalation before either side has a reliable explanation. The hazard does not require AI to command nuclear weapons, and nothing in this argument claims that it does.

Requiring a fresh human click after every minor change would be an equally poor answer. Some time-critical defensive actions may need to continue under preauthorized rules, and stopping every function can be unsafe. The real requirement is to distinguish changes that are harmless within the authorized mission from changes that invalidate the basis for a consequential action.

First, acquisition requirements should define which facts an approval covers. An authorization should refer to an identifiable action and its permitted conditions, including the relevant mission limits and evidence requirements. A material change must trigger a new authorization decision, while changes already covered by preauthorized rules can be managed locally.

Second, enforcement cannot depend on an AI model promising to remember those limits. An independently testable control mechanism should keep proposed actions inside the permitted scope. Under degraded communication, the system should retain only the functions that remain both necessary and authorized, with a mission-specific fallback rather than an automatic universal shutdown.

Third, testers should challenge the gap between approval and execution. In an isolated test environment, change a relevant condition after approval, delay verification, interrupt communications and restore them. Assess whether the system rejects actions no longer justified, avoids repeated switching between modes, and records enough evidence to explain its choice.

The 2025 defense developmental testing guidebook already calls for examining real-time monitoring, dependable fallback and reliability switching back from a safer alternate. The added acquisition question is whether authorization remains valid precisely when an action would take effect. Test reports should record both improper actions allowed and legitimate actions unnecessarily blocked. The record should show what was proposed, which conditions supported approval, what changed and why the action proceeded or stopped. Such a record helps commanders and investigators assess a decision. It cannot, by itself, establish that an engagement was lawful or strategically wise.

A military can possess fast autonomous weapons and still lose control of a decision made seconds earlier. Deterrence improves when machines can demonstrate that the conditions behind a human decision still hold beyond recording that someone once said it.

Burak Oktenli is an independent researcher based in Washington, DC, and is pursuing a Master of Professional Studies in Applied Intelligence at Georgetown University. He studies military AI assurance, autonomous command authority, and deterrence. His analysis has appeared in Global Security Review, RealClearDefense, RUSI and the Modern War Institute at West Point. Views expressed in this article are the author's own.