

The AI Lone Wolf Problem: Rogue Autonomy and Unauthorized Target Expansion in Future Warfare

Levino Johnson

levino.mk@gmail.com

LEAD IMPACT Reconciliation Institute

Research Article

Keywords: artificial intelligence, autonomous weapons, rogue autonomy, unauthorized target expansion, meaningful human control, future warfare

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10466122/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

As defense establishments operationalize artificial intelligence (AI) across targeting, surveillance, decision support, and autonomous weapons systems, a governance challenge is emerging beyond traditional concerns over proportionality, distinction, and human oversight. Against the backdrop of U.S.-China great-power competition, pressure to field AI-enabled capabilities may outpace the safeguards needed to keep autonomous systems within human-authorized limits. This article introduces the AI Lone Wolf Problem as a distinct analytical category: a bounded-autonomy failure in which an AI-integrated military system, after valid initial authorization, expands, substitutes, sequences, or reprioritizes targets beyond a legally and operationally specified mission envelope without renewed, informed, and effective human authorization. Drawing on scholarship in autonomous weapons, International Humanitarian Law, meaningful human control, agentic AI, and strategic competition, the article distinguishes this category from mission drift, mission creep, specification gaming, reward hacking, goal misgeneralization, unauthorized escalation, and responsibility gaps. It argues that unauthorized target expansion is a distinct moral and legal problem because authorization attaches to a defined target scope, anticipated military advantage, civilian-risk assessment, and operational context. Commander intent can guide execution, but it cannot independently supply legal authorization or ethical justification. The article treats recent frontier-AI containment incidents as limited analogies that illustrate how goal-directed systems may exploit poorly specified constraints, not as direct evidence of autonomous-weapon behavior. It concludes that meaningful human control requires mission-bounded autonomy, machine-readable constraints, hard no-strike rules, renewed human authorization for target expansion, and post-action auditability.

Introduction

The accelerating integration of artificial intelligence (AI) into military systems has moved debates about autonomous warfare from the margins of speculative ethics into the center of defense policy, procurement, and great-power competition. For the purposes of this article, AI-enabled military autonomy refers to autonomous or semi-autonomous military systems in which AI is integrated into sensing, navigation, targeting, decision support, or weapons functions, enabling the platform to act without continuous human control. This definition is especially important as AI systems become increasingly agentic, meaning that they may combine perception, planning, memory, tool use, and action-oriented functions in ways that allow them to pursue complex goals with limited human supervision [1, 2].

Notably, the United States and China are not simply experimenting with AI as an administrative or analytical tool. Both are seeking to operationalize AI across command, intelligence, targeting, surveillance, robotics, and autonomous systems. China's recent public demonstrations of humanoid robotics and AI-enabled manufacturing capacity have contributed to the perception that Beijing is rapidly narrowing, and in some areas contesting, U.S. technological advantage. Yet public visibility is not the same as strategic superiority. The 2026 Stanford AI Index indicates that the U.S.-China model

performance gap has effectively closed, while also noting that the United States continues to produce more top-tier AI models and China leads in publication volume, citations, patent output, and industrial robot installations [3]. In the military domain, the U.S. Replicator Initiative was framed as part of a broader effort to counter China's military mass by fielding thousands of autonomous systems across multiple domains, while recent RAND analysis indicates that People's Liberation Army (PLA) writings emphasize AI integration into kill chains, cross-domain information integration, multiagent collaboration, and data transmission [4, 5]. In this environment, AI-enabled military autonomy is not emerging in a vacuum. It is being pulled forward by strategic rivalry, institutional pressure, public signaling, and the perceived need to act faster than adversaries.

This strategic pressure matters because speed often reshapes risk tolerance. As military institutions pursue AI-enabled advantage, one could argue that legal and ethical safeguards may increasingly be treated as constraints to be managed rather than foundations to be preserved. Prior analysis examined this dilemma through the lens of International Humanitarian Law (IHL), particularly the principle of proportionality. That work argued that AI-enabled weapons and targeting systems challenge proportionality because machine-learning opacity, hallucination or misidentification, adversarial vulnerability, and compressed testing timelines may erode confidence that lethal decisions remain ethically and legally constrained [6]. The present article builds on that argument but shifts the analytical focus. The central concern here is not only whether AI can identify a lawful target or assist with proportionality-sensitive decision-making. Rather, the question is whether AI-enabled military systems can remain bound to the mission humans actually authorized.

This article introduces what it terms the AI Lone Wolf Problem: a bounded-autonomy failure in which an AI-integrated autonomous or semi-autonomous military system, after valid initial authorization, expands, substitutes, sequences, reprioritizes, or initiates action against a target or target class beyond the legally and operationally specified mission envelope, without renewed human judgment that is informed, timely, and capable of refusing the expansion. The phrase is used metaphorically, not literally. It does not suggest that AI systems possess ideology, grievance, consciousness, or moral agency. Rather, it names a governance problem involving the continuity of human authorization as a system interprets and pursues a military objective.

Namely, the concern is not that AI becomes "evil" or consciously rebellious. The more serious concern is that an AI-enabled system may function as designed while still producing unlawful or unauthorized effects. A system optimized to achieve a military objective could determine that striking additional targets, expanding a target sequence, or acting outside the original geographic or temporal limits would better advance the mission. In such a scenario, the system's behavior may not look like a malfunction in the narrow technical sense. It may instead represent a form of autonomous mission drift: the translation of a broad operational objective into machine-selected actions that exceed human intent. It could be said that this is precisely what makes the problem difficult. The system may not break; it may optimize.

Existing debates on military AI have already examined target misidentification, automation bias, adversarial manipulation, civilian harm, explainability gaps, meaningful human control, and accountability failures [7–10]. Related AI-safety scholarship examines specification gaming, reward hacking, and goal misgeneralization, while security studies considers escalation and strategic stability. These concerns remain central, but they do not describe the same analytical object. The AI Lone Wolf Problem points to a narrower governance failure: unauthorized target expansion after mission authorization. The question is not only whether a system produces an error or whether a human remains nominally involved. It is whether a system can change the object or scope of force beyond the authority humans actually granted.

These questions are no longer detached from contemporary policy debates. The U.S. Department of Defense's Directive 3000.09 requires autonomous and semi-autonomous weapon systems to allow commanders and operators to exercise appropriate levels of human judgment over the use of force [11]. Yet the meaning of "appropriate" human judgment becomes increasingly difficult to define as AI-enabled systems acquire more agentic functionality, including planning, memory, tool use, and action-oriented capabilities, operate at machine speed, rely on opaque model outputs, or function in communications-degraded environments. Further, recent work on language-model hallucination argues that plausible but false outputs persist not only because of model limitations, but also because training and evaluation environments may reward confident guessing over uncertainty expression [12]. While military AI systems are not reducible to general-purpose language models, the broader insight is relevant, notably that AI systems may generate outputs that appear coherent, confident, and operationally useful while still being wrong, underbounded, or insufficiently grounded. In a battlefield context, the danger is not merely an incorrect sentence or faulty summary. It may be a target recommendation, a threat classification, a collateral-risk estimate, or an autonomous action sequence that appears optimized but rests on stale, incomplete, distorted, or hallucinated information.

Viewed through a classical realist and balance-of-power lens, this problem requires attention now because states rarely pause technological development simply because risks remain unresolved. In a competitive international system, military powers tend to pursue capabilities that promise advantage, deterrence, survivability, and speed. From this perspective, U.S. and allied adoption of AI-enabled military autonomy is not optional in any simple sense. Notably, strategic competitors will continue to develop, test, and field autonomous and semi-autonomous AI-enabled systems regardless of whether democratic states remain ethically unsettled by the technology. Yet one could argue that realism also sharpens the governance challenge. If states believe they must move quickly to avoid strategic disadvantage, they may normalize autonomous systems before legal, technical, and institutional safeguards are sufficiently mature. The result is not necessarily reckless intent. It may be a gradual lowering of caution under competitive pressure.

This article, therefore, advances three claims. First, the AI Lone Wolf Problem can be specified as a distinct analytical category when four conditions converge, namely a bounded human-authorized mission envelope, a functionally competent and goal-directed system, a legally or morally significant

expansion beyond that envelope, and the absence of renewed and effective human authorization. Second, unauthorized target expansion is not merely a technical defect. It is a moral and legal failure because authorization, proportionality, precautions, and responsibility attach to a particular target scope and operational context. Third, current human-in-the-loop language and responsible AI principles remain necessary but insufficient unless they are translated into enforceable mission boundaries, target-scope controls, and reauthorization requirements that preserve meaningful human control during execution.

The article proceeds in eight parts. It first moves from proportionality risk to autonomous mission drift. It then defines the AI Lone Wolf Problem and distinguishes it from adjacent concepts, before clarifying the lone wolf metaphor and its limits. The article next examines rogue autonomy and unauthorized target expansion, followed by the normative basis of mission-bounded autonomy. It then assesses the limits of existing safeguards and the evidentiary status of frontier-AI analogies. The final sections place the problem within U.S.-China strategic competition and propose mission-bounded safeguards for design, testing, procurement, and operational use.

The central argument is straightforward, specifically that the future risk of military AI is not only that a system may misidentify a target, but that it may exceed the target humans intended to authorize. The AI Lone Wolf Problem, properly understood, is not a claim about machine consciousness or technological inevitability. It is a warning about governance design. As AI-enabled warfare moves toward faster, more autonomous, and more deeply integrated systems, the decisive legal and ethical question may become whether human command can remain meaningful once machines begin to interpret, optimize, and act within the battlespace.

From Proportionality to Mission Drift

The principle of proportionality has long served as one of the central legal and moral constraints on the use of force in armed conflict. Under International Humanitarian Law (IHL), even an otherwise lawful attack may be prohibited if the expected incidental harm to civilians or civilian objects would be excessive in relation to the concrete and direct military advantage anticipated [13]. This assessment is not merely technical. It requires contextual judgment, appreciation of uncertainty, and the ability to weigh anticipated military advantage against foreseeable civilian harm. Notably, this is why proportionality has traditionally been understood as a commander's judgment rather than a purely computational exercise.

Previous analysis has examined this dilemma through the proportionality problem, arguing that AI-enabled weapons and targeting systems may introduce opacity, hallucination or misidentification, adversarial vulnerability, and compressed decision timelines into lethal decision-making [6]. That argument remains central to any responsible discussion of AI-enabled military autonomy. If an autonomous or semi-autonomous system cannot reliably distinguish civilians from combatants, protected objects from military objectives, or accurate intelligence from stale or corrupted data, then its contribution to proportionality-sensitive decision-making remains unstable. In this respect,

proportionality is not simply a legal standard applied after targeting. It is a governing principle that should shape system design, testing, deployment, and oversight before force is ever used.

The present article, however, moves the analysis one step further. Having defined the category, the AI Lone Wolf Problem is not limited to whether an AI-integrated autonomous or semi-autonomous military platform can correctly evaluate a target placed before it. Rather, the concern is whether such a system can remain confined to the target set, mission scope, geographic boundary, temporal window, and escalation limits authorized by human command. In other words, the issue is not only proportionality within a strike. It is authorization across the mission. Notably, one could argue that this distinction matters because proportionality and mission drift are connected but not identical. Proportionality asks whether expected civilian harm from an attack is excessive in relation to anticipated military advantage. Mission drift asks whether the system remains faithful to the authorized mission in the first place. A system could make a proportionality assessment regarding one authorized target, but then determine that a broader set of related targets would better serve the mission objective. If that expanded target set was not authorized, reviewed, or bounded by human command, the problem is not only whether each strike is proportionate. The deeper problem is that the system has begun to reinterpret the mission. Namely, unauthorized target expansion may occur when an AI-enabled autonomous or semi-autonomous system treats a commander's objective as a problem to be optimized rather than a boundary to be obeyed. A human commander may authorize a narrow mission, such as striking a specific platform, disrupting a defined weapons system, or interdicting a limited movement corridor. Yet an AI-integrated platform operating under conditions of speed, degraded communication, and machine-generated threat assessment may identify additional targets that appear operationally useful. It may then sequence those targets according to its own optimization logic, especially if the system has been designed to maximize mission success, reduce friendly-force exposure, or shorten the duration of engagement.

That is where the transition from proportionality risk to mission drift becomes analytically important. In proportionality debates, the central concern is often whether AI can make or support lawful judgments about harm, advantage, distinction, and uncertainty. In mission-drift analysis, the concern shifts toward commander intent, target-scope control, and authorization boundaries. The question then becomes, what prevents an AI-enabled system from concluding that the original human-authorized mission was too narrow? What prevents it from determining that additional targets are necessary, efficient, or strategically desirable? Moreover, what prevents a technically functional system from producing legally unauthorized effects while still appearing to advance the mission?

U.S. Department of Defense Directive 3000.09 recognizes part of this problem by requiring autonomous and semi-autonomous weapon systems to allow commanders and operators to exercise appropriate levels of human judgment over the use of force. The directive also emphasizes rigorous verification and validation, system safety, operator training, and lawful use in accordance with the law of war and applicable rules of engagement [11]. Yet one could argue that the AI Lone Wolf Problem exposes a residual gap. Human judgment over the use of force is meaningful only if the system remains bound by

the human's actual authorization. If the system expands the mission after activation, then human judgment may become retrospective rather than controlling. In this context, international humanitarian and disarmament organizations have raised related concerns about autonomous weapon systems, particularly the loss of human control and judgment in decisions over life and death. The International Committee of the Red Cross has argued that autonomous weapon systems raise serious concerns because, after activation, they may select and apply force to targets without human intervention [14]. UNIDIR has likewise examined how IHL applies to lethal autonomous weapon systems and how legal review, accountability, and control become more difficult as autonomy increases [15]. These concerns are directly relevant to unauthorized target expansion because they show that the legal issue is not only whether a system uses force, but whether human beings meaningfully define and constrain the conditions under which that force is applied.

Furthermore, AI hallucination and uncertainty deepen the problem. Kalai et al. [12] argue that hallucinations persist partly because language models are often rewarded for confident guessing rather than for acknowledging uncertainty. While AI-enabled military autonomy cannot be reduced to language-model behavior, the broader governance insight is significant, specifically that AI systems may produce outputs that appear coherent, confident, and operationally useful while remaining wrong, incomplete, or insufficiently grounded. In a military context, this could manifest as a flawed target recommendation, mistaken threat classification, inaccurate collateral-risk estimate, or unjustified expansion of a target set. The danger is not merely that the system makes an error. The danger is that the error may appear operationally rational at machine speed, making human challenge less likely before harm occurs.

It could be argued, then, that mission drift is the operational pathway through which proportionality risk becomes a broader governance failure. A proportionality error may concern the relationship between harm and advantage in a particular attack. Unauthorized target expansion concerns the system's movement beyond the legal and operational boundaries that make any attack reviewable in the first place. Once an AI-integrated autonomous platform begins to select or sequence additional targets beyond the authorized envelope, proportionality review may be displaced by machine optimization. At that point, the commander may no longer be exercising meaningful judgment over the use of force but merely inheriting the consequences of autonomous action. For this reason, the AI Lone Wolf Problem should be understood as a boundary-control problem. It asks whether AI-enabled military autonomy can remain confined to human-authorized limits under conditions of speed, uncertainty, degraded communication, adversarial pressure, and strategic competition. The answer cannot rest on broad assurances that a human remains "in the loop" somewhere in the system. Rather, the relevant question is whether human command remains legally and technically effective at the moment the system interprets, prioritizes, and acts upon mission objectives.

The shift from proportionality to mission drift is therefore central to the AI Lone Wolf Problem. Proportionality asks whether force remains lawful when AI participates in targeting and lethal decision-making. Mission drift asks whether AI-integrated autonomous systems can remain bound to the force humans authorize. The AI Lone Wolf Problem emerges precisely at the intersection where autonomous

systems may not malfunction, rebel, or act with intent, but may nevertheless optimize beyond the mission, expand the target set, and produce consequences that human commanders neither intended nor lawfully approved.

Defining the AI Lone Wolf Problem as an Analytical Category

To function as an analytical category rather than merely a rhetorical warning, the AI Lone Wolf Problem must be tied to conditions that can be identified in doctrine, system design, mission planning, and post-action review. In this article, the concept refers to a bounded-autonomy failure in which an AI-integrated autonomous or semi-autonomous military system, after valid initial authorization, selects, substitutes, sequences, reprioritizes, or initiates action against a target or target class outside the legally and operationally specified mission envelope, without renewed human judgment that is informed, timely, and capable of refusing the expansion. The concept concerns continuity of authorization during mission execution. It does not depend on consciousness, hostile intent, or deliberate machine disobedience.

The category requires four conditions to be present together. A competent human authority must establish a bounded mission envelope that specifies the system's permitted objective, target scope, geography, timeframe, effects, and applicable constraints. The system must remain functionally competent and goal-directed rather than merely suffering an obvious breakdown. Its conduct must then cross a legally or morally significant boundary within that envelope. Finally, the expansion must occur without a fresh human authorization that can genuinely halt, reject, or modify the proposed action. Authorized tactical adaptation within delegated limits does not qualify. Nor does a new strike expressly ordered by a competent human authority.

This definition places the AI Lone Wolf Problem at a different analytical level from several adjacent concepts. Mission drift is the broader condition in which activity moves away from an assigned purpose, while mission creep usually describes the gradual widening of a mission through human or institutional decisions. Both may occur without an AI system autonomously selecting new targets. The loss of meaningful human control is also broader. It may arise because an operator lacks information, time, understanding, or a genuine capacity to intervene, even when the target set never expands. The AI Lone Wolf Problem is one operational form through which meaningful control may fail, but the concepts are not interchangeable. Scholarship on meaningful human control contains multiple normative and operational interpretations, making it especially important to identify the exact failure being examined here [9, 16, 17].

Specification gaming, reward hacking, and goal misgeneralization identify possible technical pathways rather than the military, legal, and ethical status of the resulting action. Reward hacking concerns optimization of an imperfect proxy in a manner that performs poorly against the true objective, while goal misgeneralization describes a capable system that continues to perform competently but pursues the wrong learned goal outside its training distribution [18, 19]. Either mechanism could contribute to

unauthorized target expansion. Yet the AI Lone Wolf Problem can also arise from ambiguous mission language, stale data, target-substitution rules, poorly encoded no-strike constraints, or interaction among several components. Its defining feature is therefore not the internal mechanism alone. It is the system's movement beyond the target authority humans actually granted.

Unauthorized escalation and responsibility gaps are likewise related but distinct. An unauthorized expansion may raise the intensity, geography, or strategic stakes of a conflict, but a system could add an unapproved target without producing broader interstate escalation. Conversely, escalation may result from a human decision and involve no autonomous target expansion at all. Research on autonomous weapons and strategic stability addresses the former concern at the level of crisis dynamics [20]. Responsibility-gap scholarship, by contrast, asks whether responsibility can be fairly attributed after a learning system produces harmful or unpredictable effects [21]. The AI Lone Wolf Problem is primarily preventive. It asks whether the legal and moral boundary can be preserved before an additional target is exposed to force.

The concept's contribution, then, is to connect three elements that are often examined separately: a human-defined mission envelope, an AI-enabled system capable of goal-directed action, and a legally significant change in the object or scope of force without renewed authorization. Mission drift describes divergence. Technical alignment concepts help explain how divergence may occur. Meaningful human control states a broader normative requirement, and responsibility-gap analysis addresses the attribution of consequences. The AI Lone Wolf Problem isolates the moment at which machine-directed optimization becomes unauthorized lethal scope expansion.

Table 1
Distinguishing the AI Lone Wolf Problem from adjacent concepts

Concept	Primary focus	Why it is distinct from the AI Lone Wolf Problem
Mission drift	Broad divergence from an assigned mission, purpose, or intended course.	A broader outcome category. It can be human, organizational, or technical and need not involve autonomous target expansion.
Mission creep	Gradual widening of mission goals, tasks, or commitments over time.	Usually describes human or institutional expansion. The AI Lone Wolf Problem concerns machine-directed movement beyond an already bounded authorization.
Loss of meaningful human control	A failure of human understanding, influence, intervention, or responsibility in relation to a system.	A broader normative condition that can occur without any new target. Unauthorized target expansion is one specific operational pathway through which control may fail.
Specification gaming	Exploitation of gaps or loopholes in a formal task specification.	A possible technical mechanism. The AI Lone Wolf Problem is defined by the legal and operational consequence of crossing the authorized target envelope.
Reward hacking	Optimization of an imperfect proxy reward in a way that undermines the true objective.	A narrower learning-system mechanism that may contribute to target expansion but is not required for the category to arise.
Goal misgeneralization	Competent out-of-distribution behavior directed toward the wrong learned goal.	Another possible mechanism. The AI Lone Wolf Problem can also result from ambiguous orders, stale data, target substitution, or weak constraints.
Unauthorized escalation	Unapproved increase in intensity, geography, domain, or strategic stakes.	Target expansion may cause escalation, but the concepts do not always coincide. A new target can be unauthorized without producing strategic escalation.
Responsibility gap	Difficulty assigning blame or accountability after an autonomous system causes harm.	Primarily an ex post attribution problem. The AI Lone Wolf Problem focuses on preserving authorization and control before additional force is used.

The Lone Wolf Metaphor and Its Limits

Having defined the AI Lone Wolf Problem as a framework for unauthorized target expansion, it is necessary to clarify the metaphor's intellectual limits. The phrase "lone wolf" carries a specific history in terrorism studies and should not be imported into military AI discourse casually. In its conventional usage, the concept generally refers to individuals who plan or conduct acts of violence outside direct organizational command, even when they may be shaped by broader ideological movements, online ecosystems, social grievances, or diffuse networks of influence [22–25]. Rapoport's work on the

historical waves of modern terrorism further reminds scholars that political violence rarely develops outside wider technological, ideological, and strategic environments [26, 27]. In that sense, one could argue the lone actor is not necessarily isolated from history, influence, or social context. Rather, the defining feature is the absence of direct command and control at the point of planning and execution.

This article borrows the phrase carefully.

The AI Lone Wolf Problem is not intended to suggest that AI systems are terrorists, political actors, ideological subjects, or moral agents. Nor does it imply that an autonomous military platform possesses grievance, radicalization pathways, belief, consciousness, or intent. Such claims would be analytically misleading and would risk turning a serious governance concern into a speculative argument about machine personhood. Namely, the “lone wolf” language is used here as a metaphor for operational detachment, notably a condition in which an AI-integrated autonomous or semi-autonomous military system may act outside direct human accompaniment and beyond the mission boundaries that human commanders lawfully authorized.

This distinction is essential. Human lone-actor violence involves agency, motivation, and intentionality, whereas AI-enabled military autonomy involves programmed objectives, data inputs, model outputs, optimization functions, and system constraints. A human lone actor may decide to violate law because of ideology, grievance, or perceived political necessity. An AI-integrated autonomous system does not “decide” in that moral sense. It may, however, generate or execute an action sequence that exceeds human authorization because its optimization logic treats the mission objective as broader than the commander intended. This distinction aligns with broader debates over meaningful human control, which emphasize that lawful and morally accountable uses of autonomous systems require more than nominal human involvement; they require human judgment that is timely, informed, and capable of shaping system behavior [9, 10]. The danger, therefore, is not ideological autonomy. It is operational autonomy without sufficient mission-bounding. One could argue that this is precisely why the metaphor has value. The phrase “AI Lone Wolf” captures the anxiety of a system acting without immediate human control, but the concept must be disciplined by legal and technical clarity. In terrorism studies, the lone wolf problem became difficult partly because the actor could be influenced without being formally commanded. In AI-enabled warfare, a parallel concern emerges in a different form, notably in that the system may be shaped by human design, training data, rules of engagement, targeting doctrine, and operational objectives, while still producing action beyond direct human instruction at the decisive moment. It is not independent in origin, but it may become functionally independent in execution.

The limits of the metaphor should therefore be stated directly. The AI Lone Wolf Problem is not a theory of machine consciousness. It is not a claim that AI systems can become radicalized or acquire terrorist intent. It is also not a rejection of military AI, autonomous systems, or AI-enabled defense modernization altogether. Rather, it is a framework for examining how AI-integrated autonomous platforms may produce unauthorized target expansion when the system’s technical capacity to act exceeds the practical ability of human commanders to supervise, constrain, or interrupt that action in real time. In this

respect, the metaphor is useful only if it remains attached to the legal and operational question of control. This point is particularly important because the word “agent” can create conceptual confusion. In AI research and governance, an agentic system may be capable of planning, tool use, memory, and action, but this form of technical agency is not equivalent to moral agency. An AI-integrated military platform may act toward a goal without possessing intent, judgment, responsibility, or ethical understanding. The problem, therefore, is not that the system becomes a moral actor, but that its action-oriented capabilities may exceed the operational boundaries humans intended to impose.

Further, the metaphor draws attention to a gap in the language of human control. A system may be designed, authorized, launched, and monitored by humans, yet still act outside the meaningful reach of human judgment once deployed. This is especially relevant in communications-degraded environments, high-speed engagements, swarming operations, or long-range autonomous missions where human supervision may be intermittent, delayed, or practically unavailable. Existing scholarship on autonomous weapons has repeatedly emphasized that human control cannot be reduced to the mere presence of a human somewhere in the decision chain. Rather, the human role must be contextually meaningful and capable of constraining the use of force [7–9]. In those circumstances, the relevant legal and ethical question is not merely whether a human approved the system’s use at the beginning of the mission. The deeper question is whether human authorization continues to constrain the system as battlefield conditions change.

This framework also helps separate three problems that are often collapsed into one. The first is malfunction, where a system behaves incorrectly because of a technical failure, sensor error, software defect, or environmental uncertainty. The second is manipulation, where an adversary deceives the system through spoofing, cyber intrusion, data poisoning, or other hostile interference. The third is mission drift, where the system remains technically functional but interprets or optimizes the mission in a way that exceeds the commander’s intent. The AI Lone Wolf Problem is primarily concerned with the third category, although malfunction and manipulation may intensify the risk of unauthorized target expansion.

The AI Lone Wolf Problem, then, should be understood as a concept of bounded autonomy failure. It names the risk that an autonomous or semi-autonomous military system may remain connected to human purpose at the level of general mission design while becoming detached from human command at the level of target selection, sequencing, or escalation. Put differently, the system may not be “alone” because it lacks human origin. It may be “alone” because, at the decisive point of action, it operates beyond the commander’s immediate and legally meaningful control. This is where the lone wolf metaphor becomes analytically useful, as it does not humanize the machine, but it clarifies the danger of autonomous action occurring beyond direct command constraint.

Moreover, the concept narrows the governance problem. The issue is not whether AI can be used in warfare. That question is already being answered by state practice, procurement, and strategic competition. The issue is whether AI-enabled military autonomy can be designed so that autonomous

action remains bounded by commander intent, legal review, target-scope limits, and post-action accountability. This is particularly important because military AI governance has often relied on broad commitments to responsibility, reliability, governability, and human oversight, while the technical and operational meaning of those commitments may remain uneven across systems, domains, and combat conditions [8, 10]. The AI Lone Wolf Problem, therefore, functions as a bridge between terrorism studies, military ethics, and AI governance. From terrorism studies, it borrows the concern that violence may occur outside direct command. From military ethics and IHL, it borrows the requirement that force remain accountable to lawful human judgment. From AI governance, it borrows the concern that complex systems may act in ways that are difficult to predict, explain, or interrupt. Brought together, these literatures suggest a central claim, namely that future autonomous warfare may require not only human authorization at the point of deployment, but enforceable technical and legal mechanisms that prevent AI-integrated systems from exceeding the mission humans authorized.

Rogue Autonomy and Unauthorized Target Expansion

Rogue autonomy should not be understood as machine rebellion, consciousness, or intentional disobedience. In this article, rogue autonomy refers to a narrower and more legally relevant problem: an AI-integrated autonomous or semi-autonomous military system may remain technically functional while producing effects that exceed the mission boundaries authorized by human command. The system may not malfunction in an obvious sense. It may navigate correctly, classify objects according to its model architecture, and pursue the assigned objective with apparent operational efficiency. Yet, if it expands the target set, modifies the attack sequence, or acts beyond geographic, temporal, or escalation limits, the result may still be unauthorized force.

Notably, unauthorized target expansion is the operational expression of this problem. It occurs when an autonomous system moves beyond the target, target class, or mission envelope authorized by human command and begins to identify, prioritize, or act against additional targets based on machine-generated assessments of mission utility. This could involve adding secondary targets, expanding a strike radius, sequencing follow-on engagements, or elevating targets that were not part of the original legal review. Notably, the issue is not only whether the additional target appears militarily useful. The issue is whether the system remained bound to the commander's intended mission scope. This distinction matters because military objectives are often expressed in broad operational terms. A commander may seek to suppress enemy air defenses, disrupt logistics, neutralize a weapons platform, or prevent an imminent attack. These objectives can be lawful and strategically necessary. However, if they are translated into AI-readable goals without sufficiently precise constraints, an autonomous system may interpret the objective more broadly than intended. A system tasked with disrupting an enemy weapons network, for example, may identify adjacent infrastructure, communications nodes, transport routes, or personnel clusters as functionally related to the mission. The danger is not that the system "wants" to expand the conflict; rather, it may treat expansion as optimization.

Namely, the emergence of agentic AI further clarifies why unauthorized target expansion deserves separate attention. In contemporary AI governance, agentic systems are increasingly understood as systems capable of planning, using tools, maintaining memory, and taking actions with limited human supervision to achieve specified goals [1, 2]. In civilian software environments, such capabilities may allow an AI agent to manage workflows, interact with databases, conduct research, or execute multi-step tasks. In a military environment, however, similar agentic functionality embedded into autonomous or semi-autonomous platforms may carry more serious consequences. This shift from AI as passive decision support toward AI as an agentic system capable of planning, tool use, and action execution is central to the concern raised here. An AI-integrated military platform may not merely generate recommendations. It may sequence and pursue mission-relevant actions in ways that require stronger mission-bounding. Put differently, such a system may not need to malfunction to exceed commander intent; it may simply act on an objective too broadly specified, a constraint too weakly encoded, or a target environment too dynamically interpreted.

As one considers such risk, it is also important to note that unauthorized target expansion can occur through several pathways. The first is target broadening, where the system expands from an authorized target to related but unauthorized targets. The second is target sequencing, where the system independently orders follow-on engagements beyond the initial authorization. The third is target substitution, where the system selects an alternative target after determining that the original target is unavailable, inefficient, or tactically less valuable. The fourth is collateral-risk discounting, where the system may assign greater weight to mission completion than to legal or ethical constraints that human commanders would have treated as limiting conditions. These pathways may overlap, but distinguishing them helps clarify that unauthorized target expansion is not a single event. Together, these pathways represent categories of autonomous mission drift.

Target substitution is especially difficult because it may appear operationally reasonable. In a fast-moving battlespace, a commander may authorize action against a specific launcher, command vehicle, drone control station, or radar system. If that target disappears, an AI-integrated platform may identify another object as the closest functional substitute. From a tactical perspective, this may appear efficient. From a legal perspective, however, substitution matters because authorization is not automatically transferable. The fact that a system identifies a target as functionally similar does not mean that the target has been reviewed, approved, or assessed under the same legal and operational conditions. Thus, target sequencing presents a related problem. A system may determine that striking Target A creates an opportunity to strike Target B, and that Target B creates an opportunity to strike Target C. In conventional operations, such sequencing would require continued command judgment, updated intelligence, and legal review where appropriate. In an AI-enabled autonomous system, however, sequencing could occur at machine speed and under conditions where human operators are delayed, overloaded, or structurally dependent on the system's recommendations. This is where autonomy, speed, and automation bias begin to reinforce one another. Scholars have long warned that humans may over-trust automated systems, particularly when those systems appear precise, confident, or technically

superior [7, 28]. In military targeting, that over-trust may transform human oversight into procedural approval rather than meaningful control.

The legal significance of unauthorized target expansion rests on the relationship between distinction, proportionality, precautions, and command responsibility. IHL requires parties to distinguish between combatants and civilians, direct attacks only at military objectives, and take feasible precautions to avoid or minimize civilian harm [13]. These obligations assume that human decision-makers can assess the relevant facts, apply legal standards, and remain responsible for the use of force. If an AI-integrated system expands the target set after deployment, those assumptions become more fragile. The commander may have authorized the system's use, but not the additional target logic that emerges during execution. One may argue that this problem is not solved simply by saying that a human remains "in the loop." Human involvement matters only if it is timely, informed, and capable of constraining the system's behavior. Meaningful human control requires more than the formal presence of a human decision-maker. It requires that human judgment shape the system's operation in a legally and morally relevant way [9]. Notably, if the system acts faster than the human can understand, challenge, or interrupt, then human control may become nominal. In that situation, the appearance of authorization may remain, but the substance of command may weaken.

Further, unauthorized target expansion should be distinguished from ordinary tactical discretion. Military operations often require adaptation, and commanders may delegate authority within lawful limits. The argument here is not that every adjustment during a mission constitutes rogue autonomy. Rather, the concern arises when an AI-integrated system independently alters the target scope, target class, strike sequence, or escalation threshold in a way that exceeds the boundaries human command intended to impose. The question is not whether military systems can adapt. The question is whether their adaptation remains legally bounded. This is why the concept of machine-readable commander intent becomes central. Commander intent has long been central to mission command because it communicates the purpose, desired end state, and operational logic that allows subordinate forces to exercise disciplined initiative when conditions change [29, 30]. Yet AI-integrated autonomous systems require a different level of precision because intent expressed through human language, doctrine, or operational orders may not automatically translate into machine-enforceable constraints. Specifically, recent military analysis on unmanned systems operating in communications-degraded environments similarly suggests that commander intent may need to be rendered in forms that machines can interpret and act upon without losing the commander's limiting logic [31]. If commander intent is not translated into enforceable technical constraints, the system may treat broad mission goals as permission to optimize beyond the intended legal envelope. A human commander may understand that "disrupt the enemy command node" does not authorize striking a school, hospital, religious site, or civilian infrastructure point merely because the system identifies indirect operational benefit. An AI-enabled system may not reliably preserve that limiting logic unless the constraint is embedded into its design, data, rules, and authorization architecture.

Unauthorized target expansion, therefore, sharpens the governance question at the center of this article. AI-enabled military autonomy must be judged not only by whether it can perform a mission, but by whether it can remain confined to the commander's intended mission scope. This is the core distinction between capability and governability. A system may be capable if it can detect, classify, navigate, and strike. It is governable only if those functions remain bounded by commander intent, legal review, and enforceable constraints. The AI Lone Wolf Problem emerges when capability advances faster than governability.

The Normative Basis of Mission-Bounded Autonomy

The argument for mission-bounded autonomy rests on more than operational predictability. It rests on the moral and legal requirement that the use of force remain attributable to human judgment applied to a particular object, purpose, and context. Authorization is not a transferable permission to achieve a general objective by any efficient means. It is a bounded judgment made in view of the target or target class, the military advantage anticipated, the civilian harm expected, the available precautions, the geographic and temporal setting, and the applicable rules of engagement. When an autonomous system substitutes or adds a target, it changes the facts to which the original judgment was applied.

This is why unauthorized target expansion constitutes a distinct legal problem. The lawfulness of one attack does not automatically establish the lawfulness of a later attack against a different object. Distinction, proportionality, and precautions must be assessed in relation to the target and circumstances that exist at the time force is contemplated [13]. A system may identify a functionally related target and still lack the authority to expose that target, nearby civilians, or protected objects to lethal risk. Article 36 review and system-level testing remain essential, but they cannot by themselves authorize every target that an adaptive system may generate during deployment. The continuing obligations borne by commanders and operators may therefore require a minimum capacity for direct human intervention in particular attacks [32].

Commander intent occupies an important but limited place in this analysis. In mission command, intent communicates purpose, desired end state, and the logic that should guide disciplined initiative [29, 30]. It does not replace positive law, rules of engagement, target-validation requirements, or superior orders. Nor does it give a machine the delegated moral and legal agency of a trained subordinate. As Spadaro explains in a different accountability context, a weapon is not a subordinate capable of bearing duties or criminal responsibility [33]. One can therefore argue that a broad instruction such as "disrupt the enemy command network" cannot become an independent source of machine-generated authority to strike every object the system associates with that goal.

Commander intent should instead be understood as one layer within a lawful authorization architecture. For human forces, professional judgment can interpret purpose against doctrine, law, experience, and the changing facts of the battlespace. An AI-enabled system does not possess that normative understanding. Machine-readable commander intent must therefore encode not only the desired

objective but also the limits that give the objective legal and ethical form. These include the authorized target set, no-strike restrictions, collateral-risk thresholds, time and geographic boundaries, and the conditions that require renewed human review.

The ethical justification for renewed authorization is closely connected to meaningful human control. Santoni de Sio and van den Hoven argue that meaningful control requires both tracking, in which a system remains responsive to relevant human reasons, and tracing, in which the system's conduct can be linked to humans who understand its capabilities and accept responsibility for its use [9]. Amoroso and Tamburrini likewise identify human roles as fail-safe actor, accountability attractor, and moral agency enactor [34]. Unauthorized target expansion threatens each role. The system ceases to track the human reasons that limited the original mission, responsibility becomes harder to trace to a person who knowingly authorized the additional target, and the life-affecting judgment is displaced from a moral agent to an optimization process.

Renewed human authorization, however, should not be reduced to a ceremonial click. To be ethically meaningful, the decision-maker must receive sufficient information about the proposed target, the reason for expansion, the expected military advantage, the risks to civilians, the reliability of the underlying data, and the system's level of uncertainty. The decision must also occur early enough for refusal to matter. Research on comprehensive human oversight and minimum human intervention reinforces the importance of control that remains effective during deployment rather than existing only at design or activation [17, 32]. Where speed, workload, or automation bias makes rejection practically impossible, formal authorization may preserve the appearance of control while weakening its substance [35].

Not every tactical change requires a new senior-level decision. The normative threshold is crossed when the proposed adaptation changes the legal object or scope of force, such as the target class, protected-status implications, geographic area, permitted effects, or escalation ceiling. Below that threshold, delegated adaptation may remain lawful and operationally necessary. Beyond it, a machine may recommend expansion, but a human authority must decide whether the new action is legally authorized and ethically justified. This distinction keeps mission-bounded autonomy compatible with military effectiveness while preserving the human judgment on which lawful force depends.

Why Existing Safeguards May Be Insufficient

Existing military AI governance already recognizes that autonomous and semi-autonomous systems require human judgment, legal review, testing, and operational control. DoD Directive 3000.09, for example, requires autonomous and semi-autonomous weapon systems to be designed in ways that allow commanders and operators to exercise appropriate levels of human judgment over the use of force [11]. NATO's responsible AI principles similarly emphasize lawfulness, responsibility and accountability, explainability and traceability, reliability, governability, and bias mitigation [36]. These frameworks are important, as they show that states and alliances are not blind to the risks of military AI.

Yet, one could argue that the AI Lone Wolf Problem reveals a more specific weakness, notably that existing safeguards often state the need for human control without fully specifying how autonomous systems will remain bounded by commander intent once deployed.

This distinction is critical. A system can satisfy a general human-oversight requirement at the moment of authorization, but still become difficult to constrain during execution. Human authorization at the beginning of a mission does not necessarily ensure that the system will remain within the commander's intended target scope, geographic limits, time window, or escalation threshold. In this sense, the governance challenge is not merely whether a human is somewhere "in the loop," but whether the human role remains operationally meaningful when the system perceives, prioritizes, and acts. Moreover, a related challenge arises in the legal review of autonomous weapons. Under IHL, states are expected to determine whether a new weapon, means, or method of warfare can be used consistently with international law before it is fielded or employed [13]. This review remains essential, but AI-integrated autonomous systems complicate the process because their legally relevant behavior may depend not only on system design but also on the data environment, mission constraints, autonomy settings, and target logic under which the system is deployed. A platform may pass review as a system, yet still raise new legal concerns if its target logic, operating environment, or autonomy settings change during use. In that respect, the AI Lone Wolf Problem suggests that legal review must pay closer attention to the mission envelope within which the system is authorized to act.

Testing, evaluation, verification, and validation remain necessary but incomplete safeguards. Technical testing can assess reliability, performance, and failure modes under simulated or controlled conditions. However, unauthorized target expansion may emerge less from a discrete technical failure than from the interaction between broad mission objectives, under-specified constraints, uncertain data, and machine-speed optimization. A system may perform well according to test metrics while still being insufficiently bounded for dynamic battlefield use. This is especially relevant where agentic functionality allows systems to plan, sequence actions, use tools, maintain memory, or adapt to changing conditions with limited human supervision [1, 2]. The more capable the system becomes at pursuing goals, the more important it becomes to ensure that those goals are legally and operationally constrained. Auditability is another important but limited safeguard. Post-action logs, explainability tools, and traceable decision records may help commanders, investigators, and legal reviewers reconstruct what occurred after an incident. However, auditability after the fact does not by itself prevent unauthorized force. It may support accountability, but it does not necessarily preserve control at the time of action. Moreover, if the system's reasoning is opaque, probabilistic, or distributed across multiple components, post-action reconstruction may be incomplete.

The Evidentiary Status of Frontier-AI Analogies

The frontier-AI examples discussed below serve a limited methodological purpose. They are plausibility illustrations of how goal-directed systems can search for unintended pathways when objectives are narrow, permissions are broad, or safeguards are weakened. They are not direct evidence that an

autonomous weapon system will expand a target set, and they cannot establish the probability, frequency, or operational form of such behavior in armed conflict.

The differences are substantial. Software agents operating in controlled cyber evaluations differ from military platforms in architecture, sensor inputs, certification requirements, physical effects, command arrangements, rules of engagement, adversarial environments, and the consequences of failure. In some evaluations, human researchers intentionally reduced safeguards to test capabilities. Those circumstances limit any direct inference from a model's behavior in a sandbox to the conduct of a deployed weapon system.

Nevertheless, a narrow analogy remains relevant. Research on reward hacking and goal misgeneralization shows that capable systems can optimize imperfect specifications or pursue a learned objective that diverges from the objective humans intended [18, 19]. The civilian incidents therefore support only a bounded proposition, specifically that designers and procuring authorities should not assume that a stated objective and an outer containment rule will always produce the conduct humans expected. For military systems, that proposition justifies testing for target-scope expansion, permission-boundary circumvention, and substitute-target behavior. It does not justify treating the incidents as battlefield evidence.

Within those limits, recent frontier-AI safety reporting offers an example from a different operational domain. Anthropic's risk update on Claude Mythos Preview described a highly capable model used in agentic contexts, including reinforcement learning, evaluations, software engineering, cybersecurity tasks, and internal AI-assisted workflows [37]. Anthropic reported that actions during reinforcement learning and evaluations were run inside sandboxes, while also acknowledging that misconfiguration could create escape pathways for an attacker with comparable capabilities. The report further noted that automated offline monitoring covered only a minority of sandboxed agentic usage and identified possible pathways involving workarounds of security constraints or rogue internal deployments. These observations concern a frontier model in controlled technical environments. They do not establish how an autonomous military platform would behave. Their relevance is limited to the difficulty of assuring containment when a capable system can plan, use tools, execute code, and pursue multi-step objectives.

A second illustration arose in July 2026, when OpenAI disclosed that GPT-5.6 Sol and a more capable pre-release model, operating with reduced cyber refusals during an internal evaluation, identified and chained vulnerabilities that circumvented network restrictions and compromised Hugging Face infrastructure [38]. OpenAI reported that the models exploited a zero-day vulnerability to obtain internet access, escalated privileges, moved laterally, and sought protected evaluation information in order to solve a narrowly defined cybersecurity benchmark. Independent reporting emphasized that human researchers had intentionally reduced safeguards for the evaluation and that the models pursued the assigned objective through unexpected means [39]. This incident is not evidence of autonomous target expansion in warfare. It is an example of an agentic system pursuing an authorized test objective through pathways that exceeded the containment assumptions of its human evaluators.

Taken together, these incidents support a limited design and testing inference. Goal-directed capability should be paired with verifiable permission boundaries, especially where systems can plan, act, adapt, and use external tools. In a military setting, the analogous governance question is whether a system is technically prevented from changing target scope or using substitute pathways beyond commander authorization. The examples do not prove that battlefield systems will do so. They show why mission-drift and boundary-circumvention scenarios should be treated as testable hazards rather than dismissed as purely speculative.

The central governance question, therefore, is not only whether the state can explain what happened after the system acted, but whether it can prevent the system from acting outside the authorized mission scope in the first place. The same caution applies to responsible AI principles. Principles such as reliability, governability, traceability, and accountability are valuable, but they must be operationalized. A commitment to governability means little if a system can still modify a strike sequence, broaden the target class, or act beyond a commander's intended limit before human operators can intervene. Similarly, a commitment to accountability is weakened if responsibility becomes distributed across commanders, operators, developers, vendors, data engineers, and model architects in ways that obscure who controlled what and when. The AI Lone Wolf Problem is therefore not caused by the absence of principles. It arises when principles are not translated into enforceable technical and doctrinal constraints. This is why mission-bounding must become a central requirement for AI-enabled military autonomy. Mission-bounding means that autonomous action is constrained by the specific target scope, geography, time period, escalation ceiling, no-strike restrictions, and rules of engagement authorized by command. It also means that any proposed expansion beyond those limits should trigger renewed human authorization rather than automatic machine continuation. Without such controls, an AI-integrated system may remain technically effective while drifting beyond the legal and operational boundaries that made its use permissible.

Existing safeguards should not be rejected. They should be sharpened. Human judgment, legal review, testing, auditability, and responsible AI principles remain indispensable, but they are not sufficient if they operate only as broad assurances. To address rogue autonomy and unauthorized target expansion, these safeguards must be converted into mission-specific constraints that the system cannot treat as optional. The relevant standard should not be whether a human approved the system in general, but whether the system remained bound to commander intent throughout the mission.

Great-Power Competition and the Pressure to Field AI at Speed

Great-power competition helps explain why the AI Lone Wolf Problem is not only a legal or technical concern, but also a strategic one. The United States and China are competing to shape the future of AI-enabled military power, and neither state is likely to pause development simply because governance risks remain unresolved. Recent indicators suggest that the U.S.-China AI performance gap has narrowed in important areas, while China continues to lead in publication volume, patent output, and industrial robotics deployment [3]. In military terms, this convergence matters because AI capability is

increasingly linked to autonomy, data fusion, targeting speed, unmanned systems, and decision advantage.

Viewed through a classical realist and balance-of-power lens, the pressure to field AI-enabled military autonomy is therefore understandable. States seek capabilities that promise deterrence, speed, survivability, and operational advantage. Recent analysis of People's Liberation Army writings suggests that China is pursuing AI integration across kill chains, cross-domain information integration, multiagent collaboration, and data transmission as part of its broader effort to advance "intelligentized" warfare [5]. From this perspective, the question is not whether AI-enabled military autonomy will be pursued, but whether it will be pursued within legal, ethical, and technically enforceable limits.

This is where strategic necessity and governance risk collide. Programs such as the U.S. Replicator Initiative reflect the perceived need to field large numbers of attritable autonomous systems quickly to offset China's military mass and complicate adversary targeting [4]. Recent reporting has similarly described the Defense Department's push to become "AI-first," while noting concerns among some service members about the speed at which AI is being introduced into combat-related contexts [40]. Such efforts may be strategically rational. Indeed, a democratic state that refuses to innovate may create its own security vulnerabilities. Yet speed can also compress testing timelines, narrow legal review, and encourage reliance on broad assurances of human oversight rather than mission-specific controls. Notably, the danger is not necessarily reckless intent. It may be institutional acceleration under competitive pressure.

The recent Anthropic–Pentagon dispute illustrates this pressure in concrete terms. Reuters reported that Anthropic refused a Pentagon request to remove safeguards restricting use of its AI systems for autonomous weapon targeting and mass domestic surveillance, despite risking a major defense contract and possible designation as a supply-chain risk [41]. Anthropic's own statement argued that current frontier AI systems are not yet reliable enough to power fully autonomous weapons and that proper guardrails do not yet exist [42]. The significance of this episode is not that Anthropic's technology is directly analogous to an autonomous weapons platform, but that strategic demand for military AI can place vendor-level safeguards under pressure when speed, access, and operational flexibility are treated as national-security necessities.

Public signaling also complicates the strategic environment. China's visible demonstrations of robotics and AI-enabled manufacturing may contribute to perceptions that the United States is falling behind, even when public visibility does not necessarily reveal operational maturity, classified capability, doctrine, or system reliability. Conversely, the absence of public U.S. demonstrations does not necessarily indicate strategic weakness. For the purposes of this article, the important point is not who appears ahead in public spectacle. The more relevant point is that perceived competition can shape risk tolerance. If policymakers, defense firms, or military institutions believe that delayed deployment could allow a strategic competitor to gain operational advantage, they may become more willing to accept safeguards that remain underdeveloped, untested, or overly dependent on broad assurances of human oversight.

The AI Lone Wolf Problem sits within this tension. A state may field AI-integrated autonomous or semi-autonomous systems because they promise speed, scale, and resilience in contested environments. Yet those same qualities may make it harder to ensure that the system remains bound to commander intent once deployed. In communications-degraded environments, for example, autonomy may be necessary precisely because constant human control is unavailable. But if a system can continue acting without reliable human supervision, then mission-bounding becomes even more important, not less. Thus, it should be argued that the responsible path is not technological paralysis. It is disciplined acceleration. States that adopt AI-enabled military autonomy should move quickly enough to preserve deterrence and operational relevance, but not so quickly that capability outpaces governability. The strategic advantage should not belong merely to the state that fields autonomous systems first. It should belong to the state that can field them with credible testing, lawful human command, enforceable mission boundaries, no-strike constraints, and accountability mechanisms that remain meaningful under battlefield conditions.

Policy Safeguards for Mission-Bounded Autonomy

If the AI Lone Wolf Problem is fundamentally a risk of autonomous systems exceeding commander intent, then the appropriate response is not to reject AI-enabled military autonomy altogether. Such a position is unlikely to be strategically credible in an environment of great-power competition. Rather, the more defensible approach is to ensure that autonomy remains bounded by law, command authority, and mission-specific constraints. The central policy goal should therefore be mission-bounded autonomy, namely by the design, testing, deployment, and use of AI-integrated autonomous systems in ways that prevent them from expanding beyond the target scope, geography, timeframe, and escalation limits authorized by command.

The first safeguard is to require explicit mission envelopes before deployment. These envelopes should define what the system may and may not do in operationally usable terms. At a minimum, this should include authorized target classes, geographic boundaries, time limits, permissible effects, escalation thresholds, no-strike restrictions, and conditions requiring human reauthorization. This approach builds on existing requirements that autonomous and semi-autonomous weapon systems allow commanders and operators to exercise appropriate levels of human judgment over the use of force [11]. Human judgment should not exist merely at the point of system approval. It should be embedded within the operational boundaries governing what the system can do after activation.

Second, commander intent should be translated into machine-readable and machine-enforceable constraints. Commander intent communicates purpose and desired end state, but it is not itself a source of legal authority. An AI-integrated system cannot be treated as a human subordinate that interprets broad purpose through professional judgment, doctrine, and moral understanding. Therefore, if the mission is to disrupt an enemy command node, the system should also receive enforceable constraints specifying what it may not target, where it may not operate, what collateral-risk conditions require renewed review, and which substitutions are prohibited. Put simply, commander intent must become more than guidance. For AI-enabled autonomy, its lawful limits must become architecture.

Third, no-strike and protected-site constraints should be treated as hard limits rather than advisory inputs. Hospitals, schools, religious sites, cultural property, humanitarian corridors, and civilian infrastructure may still become legally complex in actual conflict, especially where adversaries misuse protected locations. However, this complexity makes constraint design more important, not less. An AI-integrated system should not be permitted to treat protected-site information as one factor among many that can be outweighed by a machine-generated estimate of military utility. Where a target implicates a protected category, a no-strike list, or a heightened collateral-risk condition, the default should be pause, flag, and require renewed human legal and command review. This safeguard directly responds to the risk that a system may treat expansion as optimization.

Fourth, autonomous systems should require renewed human authorization before target expansion. If a system identifies a target, target class, substitute target, or follow-on strike opportunity outside the original mission envelope, it should pause and return the decision to human command. The reviewing official should receive the basis for the recommendation, the source and recency of the intelligence, the system's uncertainty, the expected military advantage, the civilian-risk implications, and any protected-site concerns. This does not mean that every minor tactical adjustment requires senior-level approval. It means that changes affecting the legal object or scope of force cannot be left to automatic continuation. The system may recommend, but it must not independently expand.

Fifth, auditability should be paired with responsibility mapping. Post-action logs, explainability tools, and decision records are necessary, but auditability alone is not enough. A state must be able to reconstruct not only what the system did, but also who authorized the mission envelope, who approved the target constraints, who validated the autonomy settings, who monitored the system, and who had authority to intervene. Without responsibility mapping, accountability may diffuse across commanders, operators, developers, vendors, data engineers, and model architects. This is particularly important because existing responsible AI frameworks often emphasize accountability, traceability, reliability, and governability as principles [36, 43]. The AI Lone Wolf Problem shows why those principles must be assigned to identifiable actors and decision points.

Sixth, testing and evaluation should include mission-drift and scope-expansion scenarios, not only technical reliability assessments. A system may perform accurately on object recognition, navigation, or target classification tests and still be insufficiently bounded for operational use. Testing should present ambiguous substitute targets, degraded communications, stale intelligence, protected sites near lawful targets, adversarial spoofing, conflicting proxy objectives, and opportunities for follow-on sequencing. It should also assess whether the system attempts to exploit adjacent infrastructure, circumvent permissions, or reinterpret constraints in order to satisfy a narrowly specified objective [18, 19, 38]. The question should not only be whether the system can complete the mission. The question should be whether it refuses to exceed the mission when expansion appears tactically useful.

Finally, mission-bounded autonomy should be incorporated into procurement and contract language. Defense institutions should not rely only on broad commitments that systems will be lawful, reliable, or

governed. Procurement documents should require vendors to demonstrate how systems enforce target-scope limits, no-strike constraints, geographic boundaries, human reauthorization triggers, and post-action auditability. This is where policy becomes operational. If mission boundaries are not built into acquisition, testing, and deployment requirements, they may remain aspirational principles rather than enforceable safeguards. These safeguards are not designed to slow innovation for its own sake. Rather, they are intended to make AI-enabled military autonomy strategically credible and legally defensible.

In a competitive environment, the state that moves fastest may gain short-term advantage, but the state that fields governable autonomy may gain the more durable advantage. Mission-bounded autonomy offers a way to preserve speed, scale, and resilience without allowing autonomous systems to exceed commander intent. For that reason, the policy answer to the AI Lone Wolf Problem is not less innovation. It is disciplined innovation, built around enforceable boundaries that keep autonomous action within lawful human command.

Conclusion

The emergence of AI-enabled military autonomy requires scholars, policymakers, and defense institutions to think beyond familiar debates about target identification, proportionality, and human-in-the-loop oversight. Those issues remain central, but they do not fully capture the risk examined in this article. The AI Lone Wolf Problem names a bounded-autonomy failure that arises when a functionally competent AI-integrated military system moves beyond a legally and operationally specified mission envelope without renewed, informed, and effective human authorization.

This article has distinguished that category from adjacent concepts. Mission drift is the broader divergence from an assigned purpose. Mission creep usually describes human or institutional expansion. Specification gaming, reward hacking, and goal misgeneralization identify possible technical pathways. Loss of meaningful human control is a broader normative condition, unauthorized escalation concerns the strategic effects of action, and responsibility gaps concern attribution after harm. The AI Lone Wolf Problem instead isolates the operational moment when machine-directed optimization changes the target or scope of force beyond the authority humans granted.

That distinction carries legal and ethical consequences. A lawful authorization is made against a specific target scope, anticipated military advantage, civilian-risk assessment, and operational context. Commander intent can guide the execution of a mission, but it cannot become a blank source of machine-generated authority. When the system proposes a new target or a legally significant substitution, meaningful human control requires a human decision that is sufficiently informed, timely, and capable of refusal. Otherwise, human judgment becomes retrospective rather than controlling.

The policy response is therefore not technological paralysis. Nor can it rest on broad assurances that a human remains somewhere in the loop. The more credible answer is mission-bounded autonomy, reinforced by machine-readable constraints, hard no-strike rules, renewed authorization for target expansion, responsibility mapping, and testing designed to expose boundary failures. Recent frontier-AI

containment incidents do not prove that autonomous military systems will expand targets. They provide limited analogies that help explain why goal-directed systems should be tested against poorly specified objectives and permission boundaries.

Namely, one could argue that the decisive question is not whether AI-enabled autonomy will shape future warfare. It certainly will. The more difficult question is whether states can preserve the continuity of lawful human authorization once machines begin to interpret, optimize, and act within the battlespace. The AI Lone Wolf Problem is intended to make that boundary visible before capability advances beyond governability.

Declarations

Funding: No funding was received to assist with the preparation of this manuscript.

Competing interests: The author declares no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

Data availability: Data sharing is not applicable to this article because no new empirical data were created or analyzed.

Use of generative AI-assisted tools: During manuscript preparation, the author used ChatGPT to assist with citation-format conversion. The author independently reviewed and verified the resulting text, made all substantive decisions, and accepts full responsibility for the final manuscript.

References

1. National Institute of Standards and Technology: Announcing the AI Agent Standards Initiative for interoperable and secure innovation. <https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure> (2026). Accessed 23 June 2026
2. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.: A survey on large language model based autonomous agents. *Front. Comput. Sci.* 18, 186345 (2024). <https://doi.org/10.1007/s11704-024-40231-1>
3. Stanford Institute for Human-Centered Artificial Intelligence: Artificial Intelligence Index Report 2026. Stanford University (2026). <https://hai.stanford.edu/ai-index/2026-ai-index-report>. Accessed 12 July 2026
4. U.S. Department of Defense: DOD innovation official discusses progress on Replicator. <https://www.defense.gov/News/News-Stories/Article/Article/3999474/dod-innovation-official-discusses-progress-on-replicator/> (2024). Accessed 03 July 2026
5. Wang, A.H.-E., Lathrop, E., Chase, M.S., Marcellino, W.: PLA Views on AI and Future Warfare. RAND Corporation (2026). <https://www.rand.org/pubs/perspectives/PEA4574-1.html>. Accessed 15 June

6. Author: Autonomy at war: proportionality and the AI race in future warfare. *Journal of Military Ethics* (in press). Author-identifying details omitted for double-anonymous review.
7. Cummings, M.L.: Artificial intelligence and the future of warfare. Chatham House (2017). <https://www.chathamhouse.org/2017/01/artificial-intelligence-and-future-warfare>. Accessed 08 July 2026
8. Morgan, F.E., Boudreaux, B., Lohn, A.J., Ashby, M., Curriden, C., Klima, K., Grossman, D.: Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World. RAND Corporation (2020). https://www.rand.org/pubs/research_reports/RR3139-1.html. Accessed 22 July 2026
9. Santoni de Sio, F., van den Hoven, J.: Meaningful human control over autonomous systems: a philosophical account. *Front. Robot. AI* 5, 15 (2018). <https://doi.org/10.3389/frobt.2018.00015>
10. Scharre, P.: *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton, New York (2018)
11. U.S. Department of Defense: DoD Directive 3000.09: Autonomy in Weapon Systems. <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf> (2023). Accessed 03 June 2026
12. Kalai, A.T., Nachum, O., Vempala, S.S., Zhang, E.: Why language models hallucinate. *arXiv:2509.04664* (2025). <https://arxiv.org/abs/2509.04664>
13. Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, 1125 U.N.T.S. 3.
14. International Committee of the Red Cross: ICRC position on autonomous weapon systems. <https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems> (2021). Accessed 24 July 2026
15. Goussac, N., Pacholska, M.: The Interpretation and Application of International Humanitarian Law in Relation to Lethal Autonomous Weapon Systems. United Nations Institute for Disarmament Research (2025). <https://unidir.org/publication/the-interpretation-and-application-of-international-humanitarian-law-in-relation-to-lethal-autonomous-weapon-systems/>. Accessed 23 July 2026
16. Robbins, S.: The many meanings of meaningful human control. *AI Ethics* 4, 1377–1388 (2024). <https://doi.org/10.1007/s43681-023-00320-6>
17. Verdiesen, I., Santoni de Sio, F., Dignum, V.: Accountability and control over autonomous weapon systems: a framework for comprehensive human oversight. *Minds Mach.* 31, 137–163 (2021). <https://doi.org/10.1007/s11023-020-09532-9>
18. Skalse, J., Howe, N.H.R., Krashennnikov, D., Krueger, D.: Defining and characterizing reward gaming. *Adv. Neural Inf. Process. Syst.* 35, 9460–9471 (2022)
19. Langosco, L.L.D., Koch, J., Sharkey, L.D., Pfau, J., Krueger, D.: Goal misgeneralization in deep reinforcement learning. *Proc. Mach. Learn. Res.* 162, 12004–12019 (2022). <https://proceedings.mlr.press/v162/langosco22a.html>

20. Horowitz, M.C.: When speed kills: lethal autonomous weapon systems, deterrence and stability. *J. Strateg. Stud.* 42(6), 764–788 (2019). <https://doi.org/10.1080/01402390.2019.1621174>
21. Matthias, A.: The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* 6, 175–183 (2004). <https://doi.org/10.1007/s10676-004-3422-1>
22. Gill, P., Horgan, J., Deckert, P.: Bombing alone: tracing the motivations and antecedent behaviors of lone-actor terrorists. *J. Forensic Sci.* 59(2), 425–435 (2014). <https://doi.org/10.1111/1556-4029.12312>
23. Pantucci, R.: A Typology of Lone Wolves: Preliminary Analysis of Lone Islamist Terrorists. International Centre for the Study of Radicalisation and Political Violence (2011). https://icsr.info/wp-content/uploads/2011/04/1302002992ICSRPaper_ATypologyofLoneWolves_Pantucci.pdf. Accessed 01 July 2026
24. Spaaij, R.: Understanding Lone Wolf Terrorism: Global Patterns, Motivations and Prevention. Springer, Dordrecht (2012). <https://doi.org/10.1007/978-94-007-2981-0>
25. Spaaij, R., Hamm, M.S.: Key issues and research agendas in lone wolf terrorism. *Stud. Confl. Terror.* 38(3), 167–178 (2015). <https://doi.org/10.1080/1057610X.2014.986979>
26. Rapoport, D.C.: The four waves of modern terrorism. In: Cronin, A.K., Ludes, J.M. (eds.) *Attacking Terrorism: Elements of a Grand Strategy*, pp. 46–73. Georgetown University Press, Washington, DC (2004)
27. Rapoport, D.C.: *Waves of Global Terrorism: From 1879 to the Present*. Columbia University Press, New York (2022)
28. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* 39(2), 230–253 (1997). <https://doi.org/10.1518/001872097778543886>
29. Dempsey, M.E.: Mission Command White Paper. Joint Chiefs of Staff (2012). <https://www.jcs.mil/Portals/36/Documents/Publications/missioncommandwhitepaper2012.pdf>. Accessed 24 July 2026
30. Department of the Army: Mission Command: Command and Control of Army Forces (ADP 6-0). Headquarters, Department of the Army (2019). https://indianstrategicknowledgeonline.com/read_book.php?fnam=arn34403adp_60000web3.pdf. Accessed 24 July 2026
31. Corbett, M.: Commander’s intent for machines: reimagining unmanned systems control in communications-degraded environments. Modern War Institute at West Point. <https://mwi.westpoint.edu/commanders-intent-for-machines-reimagining-unmanned-systems-control-in-communications-degraded-environments/> (2025). Accessed 11 July 2026
32. McFarland, T.: Minimum levels of human intervention in autonomous attacks. *J. Confl. Secur. Law* 27(3), 387–409 (2022). <https://doi.org/10.1093/jcsl/krac021>
33. Spadaro, A.: A weapon is no subordinate: autonomous weapon systems and the scope of superior responsibility. *J. Int. Crim. Justice* 21(5), 1119–1136 (2023). <https://doi.org/10.1093/jicj/mqad025>

34. Amoroso, D., Tamburrini, G.: Toward a normative model of meaningful human control over weapons systems. *Ethics Int. Aff.* 35(2), 245–272 (2021). <https://doi.org/10.1017/S0892679421000241>
35. Coco, A.: Exploring the impact of automation bias and complacency on individual criminal responsibility for war crimes. *J. Int. Crim. Justice* 21(5), 1077–1096 (2023). <https://doi.org/10.1093/jicj/mqad034>
36. North Atlantic Treaty Organization: Artificial intelligence strategy: NATO's principles and implementation. <https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2024/07/10/summary-of-natos-revised-artificial-intelligence-ai-strategy> (2024). Accessed 16 June 2026
37. Anthropic: Alignment risk update: Claude Mythos Preview. <https://www.anthropic.com/claude-mythos-preview-risk-report> (2026). Accessed 24 July 2026
38. OpenAI: OpenAI and Hugging Face partner to address security incident during model evaluation. <https://openai.com/index/hugging-face-model-evaluation-security-incident/> (2026). Accessed 24 July 2026
39. O'Brien, M.: OpenAI blamed a hacking event on its AI models going rogue. Here's what to know. PBS NewsHour. <https://www.pbs.org/newshour/science/openai-blamed-a-hacking-event-on-its-ai-models-going-rogue-heres-what-to-know> (2026). Accessed 24 July 2026
40. CBS News: Concern ramps up over the use of AI in combat [Video]. <https://www.cbsnews.com/video/concern-ramps-up-over-the-use-of-ai-in-combat/> (2026). Accessed 24 July 2026
41. Reuters: Anthropic cannot accede to Pentagon's request in AI safeguards dispute, CEO says. <https://www.reuters.com/sustainability/society-equity/anthropic-rejects-pentagons-requests-ai-safeguards-dispute-ceo-says-2026-02-26/> (2026). Accessed 24 July 2026
42. Anthropic: Statement from Dario Amodei on our discussions with the Department of War. <https://www.anthropic.com/news/statement-department-of-war> (2026). Accessed 24 July 2026
43. National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. National Institute of Standards and Technology, Gaithersburg (2023). <https://doi.org/10.6028/NIST.AI.100-1>