



The Mitchell Forum

Machine Learning & Artificial Intelligence in Military Operations: The Imperative of Human Judgment

by Maj Gen Thomas D. Taverney, USAF (Ret.)

About the Forum

The Forum presents innovative concepts and thought-provoking insight from air and space experts here in the United States and across the globe. In order to afford publishing opportunities for thoughtful perspectives, Mitchell Institute's Forum provides high visibility to writing efforts on issues spanning technology and operational concepts, defense policy and strategy, and unique interpretations of changing geopolitical relationships.

The opinions, conclusions, and recommendations expressed or implied in this article are those of the author and do not necessarily reflect the official policy or position of the U.S. Government, Department of Defense/War, the U.S. Air Force, the U.S. Space Force, or the Mitchell Institute for Aerospace Studies.

The Central Issue: Balancing AI, Autonomy, & Human Judgment ____

Machine learning (ML) and artificial intelligence (AI) offer unprecedented opportunities to improve military effectiveness, speed, and situational awareness. However, as these systems become more capable, the risks associated with autonomous decision-making increase significantly.

Autonomous operation may enable a huge speed advantage within clearly defined defensive scenarios and established Rules of Engagement. However, decisions involving lethal force, significant collateral damage, strategic escalation, or destruction of critical infrastructure should require meaningful human involvement and approval. The most responsible path forward is not to remove humans entirely from military decision-making, but to determine when it remains appropriate, necessary, or operationally relevant.

AI should serve as a powerful advisor and analytical tool; at the same time, humans should retain ultimate authority over actions that may result in loss of life, collateral damage, strategic escalation, or destruction of significant infrastructure. Human judgment remains essential for interpreting context, applying ethical principles, understanding strategic implications, and exercising accountability.

The purpose of this paper is to encourage discussion about how the United States can aggressively exploit the military advantages of AI while preserving human judgment where the consequences demand it.

Defining Machine Learning & Artificial Intelligence

Machine Learning (ML)

Machine learning is a subset of artificial intelligence that enables computer systems to learn patterns from data and use those patterns to make predictions, classifications, recommendations, or decisions without being explicitly programmed for every scenario. ML systems generally operate within defined parameters and are optimized for specific tasks using historical and real-time data. ML often makes bad assessments when used on data outside of the defined parameters. There are ways to measure incoming data and make sure it's within the reasonable purview of a model, but of course this isn't always done.

Machine Learning Applications

Machine Learning excels in environments where large volumes of data must be processed rapidly and consistently. Examples include Intelligence analysis of satellite imagery and sensor data, target detection and classification using radar, infrared, and electro-optical systems, predictive maintenance of aircraft, ships, and vehicles, cybersecurity monitoring to identify anomalies and potential intrusions, and logistics optimization for supply chains and force deployment.

In these applications, ML systems can operate without human intervention within established rules provided by the military and, when something does not fall within these preconceived rules, it can provide recommendations to the human decision maker based on historical data and learned patterns. This isn't always how ML is set up—but it should be. Any ML system should have protections to monitor data to ensure that any data outside the span of its training set is flagged for human consideration; without that, the model can actually process that data and be highly confident in its (wrong) answer.

Artificial Intelligence (AI)

Artificial intelligence refers to computer systems designed to perform tasks that typically require human intelligence, including perception, pattern recognition, planning, reasoning, language understanding, decision support, and adaptive problem solving. While the distinction between ML and AI is not always precise in technical literature, it provides a useful framework for understanding varying levels of autonomy, adaptability, and operational risk.

Advanced AI systems may generate novel courses of action, adapt to changing environments, coordinate multiple systems, and recommend solutions that were not explicitly anticipated by their designers or commanders. The level of autonomous behavior is what is key.

Artificial Intelligence Applications

Artificial Intelligence can provide substantial benefits when assisting commanders and military planners in complex decision environments. Its functions could include developing alternative operational plans, coordinating autonomous systems across multiple domains, adapting tactics to changing battlefield conditions, dynamic resource allocation, recommending courses of action under uncertainty, and supporting strategic planning and campaign analysis. These systems may generate solutions not explicitly programmed by human operators, potentially increasing military effectiveness while complicating adversary planning and decision-making.

Ethical & Operational Considerations: The Risks of Autonomous Decision-Making

Military operations differ from many commercial applications because errors can have irreversible consequences. An action that may appear optimal to an AI may be strategically disastrous. It's worth considering how certain commercial industries are using AI/ML and

putting guardrails on it, like medical devices, auto safety, or energy infrastructure. These are commercial (or civil) applications with significant implications.

The most consequential immediate potential risks resulting from the loss of human context in decision-making include collateral damage, data spoofing, and corruption. Humans can also weigh downstream considerations such as potential political consequences, the strategic environment, escalation control, and, most importantly, the ever-changing rules of engagement in a dynamic conflict.

AI learning is probabilistic, and AI cannot necessarily understand the consequences of the decisions it makes, especially unintended effects. Additionally, AI use can fall subject to such things as “model shift,” which can cause questionable decisions as real-world conditions drift and change. And, since AI accesses all data across the digital world, our adversaries could potentially insert data to corrupt U.S. decision making, otherwise known as “data poisoning.” Therefore, the U.S. military must consider what data is used for training and ongoing learning and whether it can control and assure the provenance of that data.

Complacency is another important consideration. If AI is right most of the time (99.9 percent) warfighters will get comfortable and begin to assume every decision it makes is correct. When a mission fails based on a 0.1 percent erroneous AI decision, the results could be catastrophic. How does human monitoring of AI-driven operations assess when these edge cases happen and properly seize manual control in an expeditious manner? How can warfighters recognize when the edge case happens with enough time and awareness to intervene before disaster occurs?

Two Sides to The Debate

On one side of the spectrum, machine learning (ML) and artificial intelligence (AI) have a huge positive potential to significantly

improve military operations by enabling faster decision-making, improved situational awareness, enhanced logistics, greater operational effectiveness, and increased survivability of military forces. On the other side, AI capabilities also introduce profound ethical, legal, strategic, and operational challenges. Military decisions can result in loss of life, collateral damage, destruction of critical infrastructure, or unintended escalation of conflict. Unlike many commercial applications, errors in military operations may have irreversible consequences.

Modern conflicts further complicate the application of autonomous technologies. Adversaries frequently co-locate military assets and personnel within civilian populations, near protected sites, hospitals, schools, religious facilities, and refugee populations. While an AI system may correctly identify a legitimate military objective, it may not fully appreciate the broader humanitarian, political, or strategic implications of engaging that target under rapidly changing conditions.

The debate, therefore, centers on how much authority to give to an autonomous or AI capability. At present, Defense Directive 3000.09 requires human oversight for use of autonomous weapons, full stop. However, the growing advantages of using ML and AI for military applications may require a more nuanced approach.

As military organizations increasingly adopt AI-enabled capabilities, maintaining meaningful human control over lethal and high-consequence decisions remains essential. The United States must carefully define those situations where autonomy is necessary, as well as determine the appropriate constraints.

There are also situations in which autonomous operation may be necessary and scenarios in which the AI needs to be “free-to-fire” as a quick, defensive reaction. This could entail destroying an adversary or unattributed satellite attacking U.S. satellites. Defensive

systems designed to counter incoming missiles, hypersonic weapons, drones, cyberattacks, or attacks against critical national assets, per their reactionary nature, require response times measured in seconds or milliseconds. In these circumstances, properly constrained autonomous systems may need authority to act within predefined Rules of Engagement (ROE) to protect military forces and national interests.

Trepidation over the application of AI in potentially lethal scenarios is understandable. The National Security Presidential Memorandum-11 (NSPM-11) signed by President Trump in June 2026 replaced previous guidelines to accelerate the adoption of frontier AI models across warfighting and intelligence operations. It mandates that AI be reliable, robust, steerable, and controllable. These prudent measures can be augmented by outlining a clearer decision matrix for the use of AI and ML.

Erecting a Logical Decision Framework

While AI-enabled systems can improve the speed and scale at which military forces process information, identify patterns, assess threats, and develop courses of action, their computational capability does not eliminate the need for human judgment. In some circumstances, it makes that judgment even more important.

A fundamental principle of responsible military artificial intelligence deployment is that human judgment must remain central when decisions involve potential loss of life, significant collateral damage, destruction of critical infrastructure, strategic escalation, nuclear operations, or circumstances outside the system's anticipated operating conditions. These situations require commanders and other authorized decisionmakers to consider factors that cannot be reduced solely to

Risk Category	Representative Threats	Potential Consequences	Illustrative Safeguards
Model & Data Integrity	• Data poisoning • Model manipulation • Adversarial inputs • Information operations	• Incorrect classifications or recommendations • Compromised model behavior	• Data provenance • Trusted datasets • Adversarial testing • Model monitoring • Independent verification
Operational & Environmental Uncertainty	• Sensor spoofing • Electronic warfare • Unfamiliar conditions • Changing ROE	• Performance degradation or decisions based on incomplete or outdated information	• Sensor fusion • Confidence thresholds • Restrictive fallback states • Continuous validation
Human-Machine Decision Quality	• Automation bias • Loss of context • Collateral-effects estimation • Escalation	• Excessive reliance on AI • Failure to address legal, political, humanitarian, or strategic consequences	• Explainable decision support • Competing hypotheses • Human authorization • Escalation controls
Cyber-Physical Security	• Cyber intrusions • Model exfiltration • Supply-chain compromise • Communications disruptions	• Unauthorized manipulation • Loss of capability • Exposure of sensitive information	• Zero-trust architectures • Secure execution • SBOMs • Access controls • Hardware/software attestation

Table 1: Military Critical Decision Framework: Decisions Requiring Meaningful Human Control

algorithmic calculations, including operational context, uncertainty, proportionality, legal obligations, humanitarian consequences, and broader strategic objectives.

Potential Loss of Life

AI-enabled decision-support systems can analyze large volumes of surveillance and intelligence data, identify potential threats, and generate targeting recommendations much faster than traditional processes. Autonomous or semi-autonomous systems may also perform functions associated with identifying, tracking, or engaging targets. But, when these capabilities contribute to decisions involving lethal force, human judgment is required.

Errors in training data, sensor inputs, classification models, or underlying assumptions can cause a system to incorrectly identify a civilian or civilian object as a military target. The risk is compounded by automation bias, in which operators place excessive confidence in machine-generated recommendations. Under operational pressure, human review can become little more than confirmation of an automated recommendation rather than an independent assessment. Meaningful human control, therefore, requires that personnel have sufficient information, authority, time, and understanding to challenge or reject AI-generated recommendations before lethal action is taken.

Significant Collateral Damage

AI can substantially accelerate the military “sensor-to-shooter” process by rapidly correlating intelligence, identifying potential targets, and recommending courses of action. Although this speed can create operational advantages, it can also compress the time available for commanders to evaluate civilian risk and the broader consequences of an attack.

Human judgment remains essential to the application of international humanitarian law, particularly the principles of distinction,

proportionality, and precautions in attack. Algorithms may identify patterns and estimate probable effects, but they can have difficulty interpreting ambiguous human behavior or rapidly changing circumstances. A person entering a building, for example, may be a combatant, a civilian seeking shelter, medical personnel, or someone acting under coercion. Such contextual distinctions may be difficult to derive from sensor data alone. Decisions carrying a significant risk of collateral harm, therefore, require deliberate human assessment rather than exclusive reliance on algorithmic classification.

Destruction of Critical Infrastructure

Military operations increasingly depend on AI-enabled capabilities for cyber operations, electronic warfare, targeting, and analysis of complex infrastructure networks. Actions against electrical grids, telecommunications systems, transportation networks, water systems, healthcare facilities, or other infrastructure can produce consequences far beyond the immediate military objective.

Modern infrastructure is highly interconnected. Disrupting one component may trigger cascading failures across several sectors, potentially affecting hospitals, emergency services, sanitation, food distribution, and civilian communications. These secondary and tertiary effects may persist long after an operation has concluded. Because an AI system’s optimization objective may not fully account for these broader humanitarian and strategic consequences, decisions involving critical infrastructure require human review capable of considering both immediate military advantage and foreseeable downstream effects.

Strategic Escalation Risks

When opposing forces employ increasingly automated systems, interactions can occur at speeds that leave political and military leaders with substantially less time to interpret events, communicate intentions, or pursue de-escalation.

This creates particular risk when information is ambiguous. An algorithm could interpret a defensive deployment, anomalous sensor reading, communications failure, or electronic interference as evidence of hostile action. If automated systems on opposing sides respond to one another, an initial misunderstanding could produce a rapid cycle of escalation. Human decisionmakers, therefore, remain essential for interpreting intent, considering political and strategic context, questioning uncertain information, and determining whether restraint is preferable to an immediate response.

Nuclear Operations

The requirement for human control is most consequential in nuclear operations because errors could have catastrophic and irreversible consequences. AI may offer useful capabilities in areas such as early warning, intelligence analysis, sensor fusion, and decision support, particularly as militaries confront increasingly sophisticated delivery systems and compressed warning timelines. Those same capabilities, however, introduce risks when algorithmic outputs are difficult to explain, verify, or challenge.

Complex AI models may produce assessments without providing decisionmakers with a fully understandable explanation of how those conclusions were reached. False sensor data, adversarial manipulation, software failures, or incorrect model assumptions could also generate indications of an attack that is not actually occurring. In the nuclear context, the consequences of acting on such an error are uniquely severe. AI may therefore inform nuclear decision-making, but decisions involving nuclear weapons demand the highest levels of human authority, independent verification, deliberation, and accountability.

Unplanned Operational Situations

No AI system can be trained or tested against every condition it may encounter during military operations. Unexpected weather, unfamiliar terrain, degraded sensors, communications failures, adversary deception, corrupted data, or novel tactics can move a system outside the conditions for which its performance has been validated. AI systems may become unreliable when operating outside these anticipated parameters, and the degradation may not be immediately apparent to operators. Adversaries may deliberately exploit these weaknesses through camouflage, spoofing, data poisoning, electronic warfare, or other forms of deception. Human operators, therefore, need the ability to recognize abnormal system behavior and the authority to intervene, override, restrict, or terminate AI-enabled functions when operational conditions exceed established boundaries.

Responsible Use & Human Accountability

The United States has articulated a commitment to the responsible military use of artificial intelligence and autonomous capabilities. A central element of responsible use should be the preservation of human judgment and accountability when decisions carry consequences that are lethal, irreversible, difficult to predict, or strategically significant.

Accordingly, meaningful human control should not be understood simply as having a person somewhere “in the loop.” It requires ensuring that the responsible human decisionmaker has sufficient situational awareness, adequate time for deliberation, an understanding of the AI system’s capabilities and limitations, the authority to reject its recommendation, and the ability to intervene when conditions fall outside established parameters.

The Threats: What Our Adversaries Are Doing

The advantages of AI are not exclusive to the United States: China, Russia, and North Korea are all moving rapidly to incorporate artificial intelligence into their military capabilities. Their current efforts are particularly focused on drones and drone swarms, autonomous weapons, intelligence analysis, targeting, and command-and-control systems. This competition for military advantage in AI may prove as consequential as the missile race of the 1950s and 1960s.

One of AI's most important military advantages is speed: autonomous systems can collect information, analyze threats, select courses of action, and potentially execute operations far faster than traditional decision-making processes that require humans at every stage. In warfare, that speed can provide a significant operational advantage. Therefore, it is important to understand not only how these countries are currently using AI, but also the degree of autonomy they are willing to assign to AI-enabled systems.

China

The People's Liberation Army (PLA) is investing heavily in artificial intelligence across multiple military domains. China is developing and prototyping autonomous drone systems capable of coordinating large numbers of aircraft, including swarms that could potentially conduct attacks with limited or no direct human input. The PLA is also increasingly exploring generative AI and other advanced analytical systems to process enormous volumes of intelligence and battlefield data, allowing commanders to identify threats and make tactical decisions more rapidly.¹

AI is also being incorporated into targeting and combat systems. Chinese military research includes systems capable of piloting unmanned combat vehicles, tracking naval vessels, identifying potential targets, and supporting or executing strikes.² Taken

together, these developments suggest that China views AI not simply as an analytical tool, but as a technology that could become deeply integrated throughout the military decision and engagement process.

China's approach to AI governance reflects a significant distinction between its public diplomatic position and its military development activities. At international forums such as the Shangri-La Dialogue, Chinese military officials have frequently argued that final authority over weapons should remain in human hands. Beijing has also expressed support for United Nations-centered frameworks intended to prevent a regulatory vacuum surrounding autonomous warfare.³

At the same time, China's support for binding international restrictions has generally depended on achieving international consensus on fundamental definitions and rules. In the absence of such agreements, the PLA continues to pursue increasingly autonomous drone swarms, unmanned combat systems, and AI-enabled targeting capabilities.⁴ China's highly publicized domestic AI regulations, including its AI Safety Governance Framework and Cyberspace Administration rules, are primarily directed toward civilian and commercial AI. They emphasize social stability, data protection, cybersecurity, and alignment with government requirements rather than placing comparable restrictions on PLA military development.⁵

Russia

Although Russia initially lagged behind the United States and China in foundational AI research, it has increasingly concentrated on applied and tactical military AI.⁶ Russia has also relied on dual-use technologies, including Chinese systems and Western-produced components obtained through alternative supply networks, to support the development and deployment of unmanned and AI-enabled weapons.⁷

One major area of emphasis is the automation of the military “kill chain.” Russia has extensively deployed loitering munitions, including Lancet drones, and has worked to incorporate machine-learning capabilities that allow unmanned systems to detect, track, and engage targets with progressively less human involvement.⁸ Russia has also focused heavily on command-and-control software that connects unmanned systems with artillery and other fire units. This integration allows drones to serve not simply as individual weapons or reconnaissance platforms but as components of a broader network for identifying targets and directing fires.⁹

Shaped by the ongoing war in Ukraine, Russia’s overall approach to AI in military applications is utilitarian and incremental.¹⁰ Russia’s approach to military AI governance is, therefore, comparatively permissive. Moscow has argued in international forums that existing international humanitarian law, including the Geneva Conventions, is sufficient to govern autonomous weapons and that additional AI-specific international treaties are unnecessary.¹¹ Russia also maintains that questions of military accountability and enforcement should remain largely matters of national sovereignty rather than international oversight.

Russian doctrine nevertheless formally preserves a role for human commanders.¹² AI is generally characterized as an “intelligent component” supporting human command-and-control decisions, with commanders retaining ultimate responsibility for lethal actions. Battlefield realities, however, are pushing military systems toward greater autonomy. The extensive use of loitering munitions and other unmanned systems capable of automatically detecting and tracking targets illustrates how operational pressures can reduce the practical role of human intervention once a weapon has been launched.

North Korea

North Korea is also actively incorporating artificial intelligence into its military modernization efforts, particularly in autonomous weapons, unmanned systems, cyber operations, surveillance, and intelligence analysis. Pyongyang appears to view AI as a force multiplier that can compensate for weaknesses in conventional military technology while strengthening its asymmetric warfare capabilities.¹³ North Korean state media has reported testing AI-guided tactical cruise missiles and artillery rockets incorporating automatic target-recognition and autonomous-navigation technologies.¹⁴ These capabilities could enable weapons to identify, lock onto, and track targets dynamically during flight rather than relying entirely on predetermined targeting information.

North Korea is also developing unmanned reconnaissance and attack systems. Kim Jong Un has personally supervised tests of AI-equipped reconnaissance drones and so-called “suicide attack” drones designed to locate and strike ground targets and military vehicles.¹⁵ Research attributed to North Korean institutions also indicates interest in facial recognition, voice synthesis, multi-object tracking, and related AI technologies that could strengthen battlefield surveillance, intelligence collection, and information operations.¹⁶

Cyber warfare represents another important application. North Korea’s state-sponsored cyber organizations are exploring AI as a means of improving the speed and sophistication of cyber operations as well as supporting online influence activities.¹⁷ AI could be particularly valuable to North Korea because cyber operations provide an asymmetric means of competing with technologically and economically stronger adversaries.

North Korea nevertheless faces substantial technological constraints. International export controls restrict its access to advanced GPUs and other semiconductor technologies required for large-scale AI development and training.

The country's limited electrical infrastructure also constrains its ability to operate sophisticated data centers. Despite these limitations, its recent weapons programs indicate that AI and algorithmic warfare are becoming increasingly important elements of North Korea's military modernization.

The Geopolitical Reality

China, Russia, and North Korea are treating military AI as an increasingly important element of geopolitical and military competition. A critical issue is that neither China nor Russia has established binding domestic restrictions that significantly constrain the operational use of military AI. Instead, both countries continue to emphasize rapid capability development while preserving substantial flexibility regarding how autonomous systems may ultimately be employed.

China and Russia participate in international discussions concerning AI safety and autonomous weapons, but their positions also emphasize national sovereignty and freedom of military action. North Korea faces fewer political incentives to accept meaningful international limitations and has continued to pursue AI-enabled military capabilities, despite international sanctions and technology restrictions.

The United States and China have engaged in bilateral discussions concerning AI risks, including potential safeguards against the misuse of advanced models by non-state actors.¹⁸ These discussions demonstrate that major powers recognize some of the dangers associated with increasingly capable AI systems. They do not, however, eliminate strategic incentives for states to pursue military applications that could provide an advantage over potential adversaries.

The central challenge is, therefore, not simply whether these nations are developing military AI—they clearly are—but how much

independent authority they will ultimately delegate to it. As AI becomes capable of compressing military decision cycles from minutes to seconds, nations may face increasing pressure to automate more of the process simply to keep pace with adversaries doing the same. That dynamic creates the possibility of an escalating competition in autonomy: each side may feel compelled to remove humans from portions of the decision loop if retaining them could create a potentially decisive disadvantage in speed.

The U.S. Approach: Current Policy & Guardrails

Current Military direction on the use of AI is governed by the National Security Presidential Memorandum-11 (NSPM-11) and Department of Defense Directive 3000.09, which establish rapid integration while requiring human judgment over the use of force. The rules emphasize the development of robust, reliable systems. Department of Defense Directive 3000.09 is the foundational directive governing autonomous weapon systems. It requires commanders and operators to exercise “appropriate levels of human judgment over the use of force,” and systems must be rigorously tested and verified to minimize operational failures.

“The Political Declaration on Responsible Military Use of Artificial Intelligence” is an international framework supported by the U.S. that mandates military AI capabilities have well-defined uses and are subject to testing across their life cycles to avoid unintended consequences.

The human-in-the-loop requirement in both of these governance documents means that the military strictly prohibits the use of AI to autonomously select and fire on human targets or launch nuclear weapons without human involvement. AI is instead deployed as a decision-support tool. In addition, “no-disable” provisions are intended to protect American warfighters by requiring contracts to ensure that

Critical Decision Area	Primary Risk	Why Human Judgment Is Necessary
Potential Loss of Life	Erroneous identification or engagement of combatants & civilians	Humans must assess context, uncertainty, intelligence quality, & the consequences of lethal action rather than treating an AI recommendation as determinative
Significant Collateral Damage	Disproportionate civilian casualties or property damage	Commanders must evaluate distinction, proportionality, precautions, & the broader civilian effects of an attack
Critical Infrastructure	Cascading effects on power, water, healthcare, communications, or other essential services	Human decision-makers must consider indirect & long-term consequences extending beyond the immediate military objective
Strategic Escalation	Unintended expansion or intensification of conflict	Humans must interpret adversary intent, political context, ambiguity, & opportunities for de-escalation
Nuclear Operations	Catastrophic consequences resulting from error, misinterpretation, or false warning	The magnitude & irreversibility of nuclear decisions demand the highest degree of human authority, verification, & accountability
Unplanned Operational Situations	AI behavior outside tested or anticipated conditions	Humans must recognize when system assumptions no longer apply & intervene, restrict, or terminate AI-enabled operations

Table 2: Critical Risks of AI

commercial entities do not possess the ability to disable, degrade, or materially modify deployed AI systems that the military depends on.

Critical Factors in the Use of AI: Assessing What Can Go Wrong

The use of artificial intelligence in military applications introduces risks that extend beyond the accuracy of an individual model. AI-enabled capabilities operate within a larger context of public data, sensors, communications networks, software, hardware, human operators, command structures, and operational policies. A failure or compromise within these elements can affect the quality of an AI-generated recommendation. In dynamic combat environments, minor errors contribute to operational, humanitarian, legal, or strategic harms that can have an outsized impact. The risk, on its own, does not mean that AI is inherently less reliable than human decision-making. Human decisionmakers also have cognitive biases, incomplete information, fatigue, and other limitations. Properly engineered

AI systems may help compensate for some of these weaknesses by processing information at a greater speed and scale, applying predefined criteria more consistently, and presenting alternative courses of action. The challenge is not to eliminate AI from consequential military decisions, but to design a system in which the strengths of humans and machines complement one another, and their respective weaknesses are constrained. The principal risks can be grouped into four categories: model and data integrity, operational and environmental uncertainty, human-machine interaction and decision quality, and cyber-physical security. At the same time, a single type of risk does not necessarily occur in isolation: the implications of each are interconnected.

Model & Data Integrity

AI systems derive outputs from models, data, and assumptions developed before and during deployment. Consequently, the integrity of those inputs is fundamental to operational reliability. Data poisoning,

adversarial manipulation, a compromised model, and deliberately fabricated information can influence AI outputs in ways that may not be immediately apparent to operators. Data poisoning, for example, occurs when corrupted, mislabeled, manipulated, or otherwise misleading information enters a training or adaptation pipeline. Depending on the nature and scale of the effect, data poisoning could degrade general model performance or create vulnerabilities that emerge only under specific operational conditions. Techniques may also attempt to exploit weaknesses in perception systems by manipulating physical objects, digital inputs, or sensor environments.

The existence of these techniques does not automatically mean that an adversary can reliably defeat every AI-enabled system. Their effectiveness depends on factors such as the model's architecture, adversary access, the operating environment, sensor configuration, and available defensive measures. Nevertheless, they demonstrate why AI assurance must extend beyond traditional measures of model accuracy.

Potential safeguards include maintaining trusted datasets, establishing verifiable data provenance, conducting adversarial testing, monitoring for anomalous model behavior, and independently validating consequential outputs. Where practical, cryptographic mechanisms can help establish the authenticity and integrity of software, models, and data. These measures reduce risk but should not be treated as guarantees against compromise.

Operational & Environmental Uncertainty

AI performance can degrade when operational conditions differ substantially from those represented during development, testing, and validation. Battlefield environments are particularly challenging because adversaries deliberately create uncertainty through camouflage, deception, electronic warfare, cyber operations, and manipulation of the electromagnetic spectrum.

Sensor spoofing illustrates this problem. An autonomous platform receiving misleading GPS, radar, optical, acoustic, or other sensor information may generate an internally consistent but incorrect assessment of its environment. Electronic warfare can create similar challenges by degrading communications or denying access to external processing and data sources.

A resilient architecture should therefore avoid unnecessary dependence on any single source of information. Sensor fusion can allow systems to compare information from multiple independent modalities, while physical and kinematic validation can identify observations that are inconsistent with known system capabilities or environmental constraints. Edge-deployed systems may require a sufficient local processing capability to continue operating safely when communications are degraded. Maintaining a fail-safe or restrictive operating state is equally important. In the event of a loss of communications, conflicting sensor information, declining confidence, or departure from validated operating conditions, systems should reevaluate level of autonomy needed for and appropriate to the mission. Mission requirements and the potential consequences of errors may require the system to restrict its actions, request human review, return to a predefined state, or terminate a particular function.

Human Context & Decision Quality

One of the most significant limitations of current AI systems is not simply technical accuracy but the difficulty of representing the full context surrounding a military decision. An AI system may identify patterns in sensor and intelligence data, but it has a limited ability to account for cultural, political, legal, humanitarian, or strategic factors that were not adequately represented in its data, objectives, or operating constraints.

An urban reconnaissance system, for example, might detect armed individuals without being able to determine confidently whether they represent hostile forces, local security personnel, civilians protecting a community, or another context-dependent category. The appropriate response may depend on information that extends well beyond visual classification. Technical systems can help reduce this uncertainty through additional contextual data, confidence measures, multi-source intelligence, and decision-support interfaces. However, these measures should be understood as mechanisms for improving human-machine decision-making, not as substitutes for human judgment.

Explainable AI can be useful in this regard. It provides operators with understandable information about the evidence, assumptions, confidence, and competing factors underlying a recommendation. Rather than attempting to display the “exact reasoning” of a complex model—which may not always be technically possible or meaningful—interfaces should provide operators with enough information to assess why a recommendation was generated, what evidence supports it, and what uncertainties remain.

Collateral effects and infrastructure. AI-enabled systems may assist military planners in estimating blast effects, civilian presence, infrastructure dependencies, or other consequences of a proposed action. However, these estimates are necessarily based on available data and modeling assumptions. The challenge becomes particularly significant when targets are located near or integrated with civilian infrastructure. Damage to electrical systems, communications networks, transportation systems, industrial facilities, or water infrastructure may create second- and third-order consequences extending beyond the immediate target. These effects may be difficult to predict completely in advance.

Physics-based modeling, infrastructure dependency analysis, updated intelligence, and automated abort criteria may help reduce these risks. Nevertheless, technological safeguards should support rather than replace the commander’s obligation to consider distinction, proportionality, military necessity, precautions, and foreseeable civilian harm.

Political and strategic escalation. Military actions do not occur exclusively within a tactical environment. The use of an autonomous or AI-enabled capability can have diplomatic, political, and strategic consequences that extend beyond whether a particular engagement was technically successful.

AI-enabled operations may also compress decision timelines. When opposing forces employ increasingly automated systems, ambiguous actions could be assessed and answered before political or military leaders can fully evaluate their significance. A defensive maneuver, sensor anomaly, or communications disruption could read as a hostile activity and lead to unintended escalation.

The appropriate response is not to impose delays on every AI-enabled action. In some defensive scenarios, speed is essential. Instead, system designers and policymakers should identify escalation-sensitive decision points at which additional authorization, verification, or deliberate human review are required. Systems can also be designed with tiered authorities that distinguish routine defensive responses from actions that substantially increase the intensity or scope of a conflict.

Rules of engagement & changing operational authorities. Rules of engagement (ROE) and operational authorities can change rapidly in response to diplomatic agreements, ceasefires, changes in mission objectives, or evolving battlefield conditions. An AI-enabled system operating with outdated information

could take an action that was permitted when the mission began but is no longer authorized. This risk is particularly important for systems operating in communications-denied or degraded environments. Designers should establish clear rules governing what authority remains valid when a system can no longer confirm that its ROE, targeting data, or mission parameters are current.

One approach is to establish time-limited authorities or restrictive fallback states. If required updates or authentication cannot be obtained within a specified period, the system could revert to a more constrained mode. In such cases, an AI-governed system should not assume that its previous authorization remains valid indefinitely. For high-consequence actions, independent verification and explicit human authorization may provide additional safeguards.

Automation bias & human-machine learning. Automation bias presents a different type of vulnerability because the AI system itself may be functioning as designed. The failure occurs when humans place excessive confidence in AI recommendations.

As AI systems demonstrate high levels of reliability during routine operations, operators may become less likely to independently verify their outputs. This creates a paradox: the more reliable an automated system becomes, the easier it may be for humans to overlook the rare circumstances in which it is wrong. Those rare failures become particularly important when the consequences of a single error are severe.

Neither humans nor AI systems can achieve perfect decision-making; Establishing governance over AI use should not assume otherwise. Instead, human-machine teams should be designed so that one can identify or compensate for weaknesses in the other.

Interfaces can support this goal by displaying uncertainty, identifying conflicting evidence, and presenting competing hypotheses. Every recommendation should

not be communicated as a definitive answer. Training should also expose operators to realistic system failures, so they remain aware and capable of recognizing anomalous behavior. For particularly consequential decisions, independent human verification or multiple authorization requirements may be appropriate.

Cybersecurity Model Protection & Supply Chain Risk

Whereas AI-enabled military capabilities inherit many of the same cybersecurity risks associated with conventional software, there are additional concerns associated with model weights, training data, inference pipelines, sensors, and specialized hardware. Cyber intrusions could modify system parameters, interfere with data flows, or otherwise expose sensitive models and information. Model exfiltration is another risk, which would allow an adversary to study a system's behavior and search for exploitable weaknesses.

Supply-chain compromise introduces another potential attack path. Military AI systems may incorporate commercial software, open-source libraries, third-party components, and globally manufactured hardware vulnerable to corruption. Mitigation therefore requires a defense-in-depth approach rather than reliance on a single technical control. Measures may include trusted execution environments, strong identity and access management, cryptographic signing, software bills of materials, secure development practices, hardware and software attestation, vulnerability management, and rigorous regression and adversarial testing before deployment.

These controls can substantially reduce exposure, but no architecture should be characterized as completely tamper-proof or immune to compromise. Continuous monitoring, incident response, recovery mechanisms, and the ability to operate safely in a degraded state remain essential.

Designing for Failure, Not Assuming Perfection

The central challenge in military AI assurance is not achieving a hypothetical state in which an AI system is always correct. Complex military environments make such a standard unrealistic for humans and machines alike. Instead, systems should be engineered with the assumption that models will sometimes be wrong, sensors will sometimes fail, communications will sometimes be disrupted, adversaries will attempt deception and compromise, and humans will sometimes misinterpret automated recommendations.

Responsible military AI architecture should therefore focus on preventing individual failures from becoming catastrophic failures. This requires layered technical safeguards, clearly defined operational boundaries, continuous validation, meaningful human oversight, and mechanisms that allow authorized personnel to question, constrain, override, or terminate AI-enabled actions when necessary. The measure of a trustworthy military AI system is consequently not whether it eliminates uncertainty, but

whether it can recognize, communicate, and safely manage uncertainty while preserving human accountability for the decisions whose consequences demand it.

Translating the Principle into an Operational Model

Minimizing uncertainty and risk comes down to determining when, how often, and to what degree human operators must exercise oversight of AI-enabled capabilities and balancing this manpower requirement with operational goals for speed and performance. Some missions will require a far higher degree of caution to mitigate potential outcomes, and others will require assuming more potential risks to accomplish the same. A way of describing these different levels of control is having a human in the loop or a human on the loop.

Human-in-the-loop (HITL). Under a human-in-the-loop framework, the AI system analyzes available data and generates recommendations. Human operators then review the recommendation and assess the relevant legal, ethical, and operational considerations. Humans make the final

Level	Human Role	AI Authority	Example applications	Governance Requirement
Decision Support	Human makes all decisions	AI provides analysis & recommendations	- Intelligence analysis - Logistics planning - Predictive maintenance	Routine oversight
Human-in-the-Loop	Human approves actions before execution	AI recommends actions	- Target recommendations - Operational planning - Cyber response options	Mandatory human approval
Human-on-the-Loop ML-like implementation	Human supervises & may intervene	AI executes within predefined parameters	- Missile defense - Counter-UAS systems - Satellite self-defense	Continuous monitoring
Conditional Autonomy	Human establishes objectives & constraints	AI executes mission within approved ROE	- Swarm coordination - Autonomous ISR collection - Contested logistics	Periodic review
Totally Autonomous	No Human intervention or participation	AI acts independently	Fully autonomous lethal	Not recommended
Hybrid of all levels	Variable: primary in mission planning & high consequence actions	AI is autonomous when following ROE's & defensive measure	Variable by application	Hybrid of human out-of-the-loop & human-in-the-loop

Table 3: Decision Framework Definitions

execution decision. The AI provides the human with information and tradeoffs that inform the decision, but it does not execute the action unless directed. Only after the human operator physically directs the action is it then executed, either by AI or authorized personnel.

Human-on-the-loop (HOTL). Under a human-on-the-loop framework, AI executes actions autonomously within predefined parameters while human operators monitor system performance. Humans retain the authority to intervene or terminate operations. This approach may be appropriate for highly time-sensitive defensive scenarios where reaction times exceed human capabilities.

Autonomy does not eliminate human authority; it changes when that authority is exercised. In some missions, the human makes the decision immediately before execution. In others, commanders establish the authority, limits, and rules beforehand and allow the system to act within those boundaries. A fundamental principle for responsible military AI deployment is the requirement for a human in the loop whenever decisions involve the potential for loss of life, significant collateral damage, the destruction of major infrastructure, or strategic escalation risks. Human operators must also be able to exercise judgment and oversight when the AI encounters unplanned operational situations.

Table 4 highlights the human role, AI role (authority), and potential governance requirements across a spectrum of human-in-the-loop versus a human-on-the-loop approaches.

No single approach is best for every possible situation. Building capabilities and operational concepts based on a clearly defined risk-benefit matrix allows for speed when executing within the defined ROEs and preserves the backstop of mandatory human intervention for high-consequence actions involving lives or critical infrastructure. At the same time, human-in-the-loop control should remain the baseline. The practical answer is a hybrid model: the degree of human control changes with the mission, the operating environment, the speed required, and the consequences of getting the decision wrong.

Recommended Governance Model

The Spectrum of Military AI Autonomy illustrates that different mission sets require different levels of human oversight. The objective is not to maximize autonomy, but rather to align the level of autonomy with the potential operational, ethical, legal, and strategic consequences of the decision being made. The above framework allows military organizations to realize operational benefits while preserving human judgment where consequences are greatest. Low-risk

Level of Autonomy	Definition	Challenges & Benefits
ML	Highly scripted, well understood processes implemented in ML S/W.	• Lacks flexibility • Only useful for well known & understood processes & actions • Check-list operations replace manuals with computer tools that interact with the user in real time
AI-Constrained	AI-provided COAs & can only follow designated ROEs. Unknown/unplanned activities requiring decisions will revert to human with decision support tools to accelerate decision making, fuse data across domains, & present option trades.	• Eliminates risks to non-combatants & critical infrastructure • Notifies human with issues & provides options • Speeds decision making with decision support trades & tools
AI	AI given freedom to make its own decisions.	• Assumes potentially lethal & escalation risks

Table 4: Spectrum of Military Autonomy

applications should undergo periodic review; moderate-risk applications should have a human-on-the-loop approach; high-risk applications should have a human-in-the-loop approach; and applications involving critical decisions on lethal force, infrastructure attacks, offensive satellite operations, or strategic escalation must require human authorization. This approach allows military organizations to gain efficiency benefits from AI while preserving human judgment where consequences are greatest.

Unplanned & novel situations. One of the greatest challenges for AI is responding to unexpected situations and novel adversary tactics. While AI can be very good at adapting to certain kinds of environmental changes and executing alternative actions, those actions, though theoretically optimal, may not fit withing current rules of engagement or account for unexpected civilian movement into an engagement area.

When uncertainty exceeds predefined rules of engagement thresholds, AI systems should automatically flag the situation and escalate to human decisionmakers. They can provide and ranked recommended actions along with a confidence assessment and an explanation for the basis of those

recommendations. The system should be programmed to always default toward caution rather than autonomous action.

Governance gates. Before an AI-enabled capability advances to a higher level of autonomy, organizations should evaluate its compliance with rules of engagement (ROE), the Law of Armed Conflict, and international humanitarian law. The evaluation should also consider explainability, traceability, reliability under contested conditions, resistance to cyberattacks and deception, the effectiveness of human intervention mechanisms, strategic escalation risks, and potential collateral effects. As a general rule, increasing operational consequences should require increasing levels of human judgment and accountability.

Ensuring AI Resilience in Contested & Degraded Environments

As the military maxim states, no plan survives first contact with the enemy. This reality remains unchanged in the era of Artificial Intelligence. Warfare is inherently chaotic, and military AI systems must operate reliably within highly contested and degraded environments.

This principle is especially critical for the U.S. Space Force, where degraded operations are not an edge case—they are the standard operating environment. Conduction missions in high-stakes environments—such as autonomous satellite operations—requires an engineering philosophy known as fail-safe bounded automation. A fail-safe or bounded control strategy restricts an automated or AI-driven system within predefined operational limits and ensures it transitions to a safe state during a failure rather than causing cascading damage. This means that when an AI system faces high operational uncertainty, it must not guess. Instead, it must systematically shed non-essential functions, preserve its physical integrity, and present structured, actionable choices to human operators.

Decision Type	ML/AI Autonomy Level
Sensor fusion & threat Detection	Strong autonomy
Defensive cyber response	Strong autonomy
Target recommendation	Moderate autonomy
Lethal engagement decision	Human-in-the-loop mandatory
Strategic infrastructure attacks	Human-in-the-loop mandatory
Unplanned situations	Human-in-the-loop mandatory

Table 5: An Example Risk-based Governance Model

Many fail-safe measures already exist. To prevent an AI from executing catastrophic actions based on uncertain or corrupted data, systems employ epistemic uncertainty estimation, a mathematic measure of how well the AI “knows what it does not know.” Out-of-distribution (OOD) detection can also enable the AI to continuously monitor environmental inputs. If sensors encounter a phenomenon outside the system’s training distribution, such as an unprecedented solar glare or a novel electronic jamming pattern, the OOD monitor flags the data as untrusted. As another failsafe example, dynamic confidence thresholding ties system actions directly to mathematical confidence scores. If confidence drops below a strict threshold, such as 95 percent target identification certainty, the system triggers an immediate algorithmic rollback. A “safe-hold” architecture can provide a deterministic, non-AI fallback state when an uncertainty threshold is breached. For orbital assets, this may involve shutting down payload cameras, pointing solar arrays at the sun to maintain power, and entering a predictable, slow-rotation “sun-safe” orbit. For tactical edge assets, the system may pause physical movement, disable weapon authorization logic, and switch to passive, low-emission reconnaissance.

Importantly, preserving human oversight in these high-stake circumstances is feasible, but only by abandoning the illusion of real-time control in contested or high-speed environments. Real-time HITL frameworks may fail during communications blackouts or hypersonic encounters. Human oversight must instead be re-engineered around three distinct pillars: triage alerts, time-bounded leases, and explainable logs. Triage alerts address the difficulty humans face in processing raw, high-bandwidth sensor data over a degraded link. The edge AI must compress complex situations into binary or multiple-choice decisions. Instead of transmitting raw data streams, the system can uplink a low-bandwidth text alert such as, “Detected: unknown object matching 70 percent threat profile. Action: divert

orbit or maintain course. Defaulting to maintain course in 120 seconds.” Time-bounded leases allow ground stations to grant the AI a limited period of autonomy. For example, a satellite may be authorized to perform autonomous station-keeping maneuvers for exactly 12 hours. If it does not receive a cryptographic “heartbeat” handshake from a human operator before the lease expires, it freezes autonomous maneuvers and awaits instruction. Finally, explainable logs enable humans to retain macroscopic oversight through post-mission forensic audits. The AI must log its decision trees using deterministic state machines alongside neural network outputs, allowing human engineers to review the exact variables that triggered a safe-state execution.

The theme of these measures is to provide human operators a greater degree of understanding and predictability about how AI systems will act. This also increases operator trust in the systems, even as they degrade. When a system degrades haphazardly, it amplifies uncertainty and can cause chaotic, cascading failures. To protect mission assurance, platforms must be engineered to degrade predictably, transparently, and monotonically. Systems should step functions downward systematically rather than fluctuate erratically.

Functional decoupling, or compartmentalization, is one example of this principle. It requires the AI engine to remain isolated from core flight or survival systems. If advanced AI targeting or data-gathering software crashes, the underlying guidance, navigation, and control code can continue to operate on a separate, hardcoded processor. No-AI deterministic fallbacks provide another layer of protection. If the AI model’s internal sanity checks fail—for example, it detects that its software weights have drifted or been manipulated—then the vehicle completely deactivates the machine-learning stack. It then reverts to legacy deterministic code, including static algorithms, lookup tables, and hardcoded geometric rules, which are less capable but fully predictable.

Principle	Definition
Meaningful human control	Humans must retain authority over decisions involving lethal force & significant destruction
Explainability	Military personnel should understand why was a recommendation made, what data supported it, what uncertainties exist, & what black-box recommendations are insufficient for high-consequence decisions
Accountability	• AI systems cannot be held legally or morally accountable • Responsibility for military actions must remain with human commanders & authorized decision-makers
Proportionality & compliance	Decisions must comply with International Humanitarian Law, the Law of Armed Conflict, ROE, & other national military policies
Positive control & witting escalation	When confidence is low or consequences are severe, systems should defer to human review
Continuous validation	Models must be continuously evaluated for bias, drift, reliability, adversarial manipulation, & operational effectiveness.

Table 6: Example of Graceful Degradation

The Core Paradigm: Cyber Resilience vs. Traditional Cybersecurity

Traditional cybersecurity focuses on perimeter defense—building a wall to keep attackers out. In a contested military AI environment, this approach is insufficient. True cyber resilience assumes the adversary has already breached the perimeter and entered the system.

The baseline requirement for military AI is a specialized zero trust architecture engineered explicitly for data and algorithms. Implementing this architecture requires a continuous, engineering-based loop of four core capabilities: anticipate, withstand, recover, and adapt.

Anticipate: Design for Controlled Failure

Resilient architecture maps out every system dependency and failure point before deployment. This means assuming compromise. Every software component should be designed under the assumption that adjacent components may already be controlled by an adversary. Attack graph mapping can also be used to identify the mathematical paths an attacker could take through the system—from data ingestion to edge deployment—to systematically

identify and eliminate single points of failure. Degraded-mode planning can likewise establish a minimum baseline of operational capability that can be maintained when advanced functions are lost. For example, if an autonomous platform loses its AI vision module to an electronic warfare attack, it must possess a hardcoded fallback that enables it to return home using basic inertial navigation.

Withstand: Contain the Damage

When a breach occurs, the system must automatically compartmentalize the threat to prevent lateral movement across the network. Micro-segmentation is a method that isolates workloads into independent micro-perimeters so that an attack on the data pre-processing pipeline is physically and logically incapable of spreading to the live inference engine. Dynamic zero trust practices can enforce continuous verification by requiring every data payload, model update, and user request to be authenticated and authorized at every step using short-lived, ephemeral cryptographic tokens. Moving target defense (MTD) is another cybersecurity strategy that dynamically shifts the system's internal configurations.

By continuously rotating encryption keys, shuffling memory layouts, and changing IP addresses, the system can disorient attackers and disrupt automated exploits.

Recover: Restore to a Known Good State

System resilience is measured by the speed at which an infrastructure can purge an infection and resume normal mission operations. An immutable infrastructure model supports recovery by deploying system components as stateless containers. If a server displays anomalous behavior, the automation layer immediately terminates the affected container and redeploys a pristine copy from a cryptographically signed, read-only registry.

Adapt: Evolve Faster Than the Threat

A resilient system learns from every close call and actively hardens itself against future iterations of an attack. Automated red teaming processes deploy continuous, automated adversarial AI systems that probe internal networks around the clock. This application of chaos engineering helps discover vulnerabilities before human adversaries can exploit them. Telemetry and feedback loops feed real-time data from edge devices back into a centralized security operations center. This information can then be used to update defensive filtering algorithms across the operational fleet within minutes.

Manage: Overcome Loss of Communications

A total loss of communications (comms-out) caused by electronic warfare, jamming, or severed satellite links requires a shift from a centralized, cloud-reliant architecture to a completely decentralized, autonomous edge framework. When the network goes dark, the governing concept should be “Mission Command for Algorithms.” Commanders establish the intent, authorities, boundaries, and prohibited actions in advance; the AI determines how to execute within those boundaries when

communications are unavailable. This protocol empowers edge systems to make local decisions aligned with the commander’s intent without phone-home capabilities.

Tactical Edge Framework: Terrestrial & Aerial Assets

Edge-native computing (no-cloud dependency). Traditional AI models rely on massive, cloud-based data centers to process information. To survive comms-out scenarios, systems must execute high-level processing locally on the tactical edge. Model quantization and pruning are optimization techniques that compress large AI models to a fraction of their original size, allowing them to run locally on ruggedized, low-power microchips, such as custom Application-Specific Integrated Circuits (ASICs), or system-on-modules, such as those embedded directly in drones, vehicles, or soldier kits. Developing technologies with an on-device inference ability enables all computer vision, sensor fusion, and threat evaluation to be performed locally on the hardware. This allows the platform to remain completely self-contained, requiring no external data packets to identify and track targets.

Inertial & alternative navigation (alternative PNT). Adversaries heavily target GPS and GNSS signals using electronic jamming and spoofing, potentially rendering standard navigation ineffective. Inertial navigation systems utilize highly precise ring laser gyros and accelerometers to calculate a vehicle’s position, orientation, and velocity through dead reckoning without relying on external signals. Visual odometry and terrain matching navigation processes employ optical cameras and AI to scan the terrain below and compare real-time features against pre-loaded, onboard 3D topographic maps, allowing the platform to determine its location with zero electronic emissions. Alternative PNT systems can further verify position by leveraging

Earth's magnetic anomalies or celestial navigation algorithms, effectively using magnetic fields and stars as un-jammable reference points.

Decentralized swarming & ad-hoc mesh networking. When long-range communications to headquarters are severed, local assets must still coordinate with one another using short-range, resilient network architectures. Mobile ad-hoc networks utilize directional, low-probability-of-intercept/low-probability-of-detection (LPI/LPD) radio frequencies to establish instantaneous, self-healing mesh networks among local drones or vehicles. If one node drops from the network, the remaining nodes automatically reconfigure their connections. Distributed consensus algorithms is a process that uses decentralized protocols, such as swarm intelligence, to allow local autonomous platforms to coordinate target assignment, fuel allocation, and mission priorities without relying on a centralized command server. Passive coordination is another process through which autonomous systems predict the actions of other platforms based on visual observation and predefined mission rules instead of actively transmitting coordination data, enabling a swarm to maintain formation and execute coordinated tactics while maintaining radio silence.

Semantic comms & delay-tolerant networking (DTN). If communications are severely throttled, intermittent, or jammed rather than completely unavailable, standard data transmission methods may become ineffective. Semantic communication models compress information into a text-like "semantic" summary rather than transmitting high-definition video or raw sensor feeds. For example, transmitting the text message "3x enemy tanks at coordinates X, Y moving North" requires only a fraction of the bandwidth needed to transmit a 50 MB video file. Store-and-forward delay-tolerant networks (DTN) employ protocols that preserve data packets when a connection is interrupted. The system stores the information locally and

subsequently forwards, or "hops," it on to the next available asset, such as a physical courier drone that returns to a friendly communications zone.

Algorithmic delegation & human-in-the-loop thresholds. The most critical challenge of a comms-out environment is the legal and ethical dilemma associated with target engagement. If a human operator cannot be reached, the system must operate within strict, predefined operational boundaries. A pre-delegated authority framework would establish clear algorithmic rules before launch. For example, if communications are lost, the AI may be authorized to use lethal force only against highly specific, pre-validated targets, such as active radar installations within a defined geographic area, while reverting to purely defensive or reconnaissance modes in all other scenarios. Fail-safe abort profiles are another fallback method to establish strict, immutable timeouts. If the system does not re-establish a secure digital handshake or clear a specified checkpoint within a designated period, it automatically sanitizes its memory banks of classified data and returns to a pre-designated recovery point.

Space Domain Framework: Orbital Assets

When a satellite experiences lost or degraded communications, standard real-time satellite operations break down. Operators cannot use a traditional HITL framework because they cannot send immediate commands or receive live telemetry. Instead, the paradigm shifts to a HOTL or even a human-in-the-past/future-of-the-loop model. The human's role transitions from a real-time pilot to a boundary setter, intent programmer, and post-facto auditor.

Pre-Mission Phase: Human-in-the-Past (The Safe Operating Envelope)

Before a satellite launches or enters a known communications blind spot, such as an orbital dead zone or intense adversarial jamming, humans can define the exact

Operational State	System Status	Algorithmic Layer
Fully mission capable	Nominal operations	Full AI & machine learning stack active
Degraded reconnaissance	Low-bandwidth/high uncertainty	Automated data filtering & compression only
Safe-hold/survival	Critical alert state	Non-AI deterministic fallback code

Table 7: Challenges of autonomy levels

boundaries of autonomous behavior. Intent-based scripting is one method that establishes high-level intent rather than relying on step-by-step commands. For example, the system could be configured to “maximize imagery collection over coordinates X while maintaining a minimum 30 percent battery reserve” rather than being directed to “turn camera 5 degrees left at 12:00.” Developing hardcoded contingency matrices can help establish immutable “if-then-else” rules within the flight software. The onboard AI determines how best to achieve the mission objective, while humans establish safety boundaries that cannot be crossed. For example, “If thruster temperature exceeds X°C, then the system must immediately shut down and enter a spin-stabilized safe mode regardless of mission priorities.”

Contested/Degraded Phase: Human-on-the-Loop (The Asynchronous Supervisor)

When communications are degraded by high latency, low bandwidth, or intermittent packets, the AI handles second-by-second spacecraft survival while the human acts as a macroscopic editor. Semantic telemetry triaging is a practice that uses the onboard AI to generate small, compressed “state vectors” save-point. This saves the system from needing to transmit megabytes of raw health and status data over a constrained connection. The human operator can review a single text line summarizing the satellite’s health and approve or veto the AI’s proposed course of action through a low-bandwidth uplink.

Assigning a “check-in” cadence allows the satellite to autonomously perform a task, such as maneuvering to avoid space debris identified by its onboard orbit propagator, while requiring it to pause at a pre-designated “gate.” Before executing the final burn, the system waits for a simple binary handshake, such as “proceed” or “abort,” from a human ground station, thereby accommodating communications delays.

Total Loss of Comms: Human-out-of-the-Loop (Bounded Autonomy)

If the link is completely severed, the human is temporarily removed from the active decision cycle. The onboard AI takes control, but its authority remains tightly bounded by human-designed protocols. Autonomous fault detection, isolation, and recovery (FDIR) is a cyber security approach that asks the AI to isolate damaged memory sectors if, for example, a radiation event corrupts the primary flight computer. The system can then switch to a redundant cold-spare processor and attempt to reorient its high-gain antenna toward Earth to re-establish the human link. Passive ephemeris maintenance is a practice that uses onboard optical sensors to track stars and Earth’s horizon through autonomous navigation, or AutoNav, without relying on ground tracking stations to update orbital positioning data. The satellite calculates its own position to prevent it from drifting off course while communications remain unavailable. Proportional self-protection authorities permit automated defensive maneuvers or sensor-shielding protocols when the AI detects an

immediate physical or cyber threat, such as an uncooperative closing satellite or localized electronic jamming, to preserve the asset until communications are restored.

Conclusion

AI must co-exist with American moral and ethical values. Warfighters should never totally submit their hearts and minds to technology. AI has tremendous advantages in leveraging human intelligence, but only with strict guardrails. Maintaining meaningful human control over unplanned critical decisions ensures that military operations remain aligned with ethical principles, legal obligations, and strategic objectives. This is why the Department of War has established five ethical principles for artificial intelligence: it must be responsible, equitable, traceable, reliable, and governable.

To operationalize these principles, the Department has published the Responsible AI Strategy and Implementation Pathway and

established the Chief Digital and Artificial Intelligence Office (CDAO) Responsible AI Toolkit. Military organizations should also leverage the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF), which provides a structured approach for identifying, assessing, and managing AI-related risks.

At the same time, adversaries are growing and maturing their AI capabilities. Machine-speed autonomy is increasingly a necessity for peer conflicts. The U.S. military should exploit AI aggressively for its enormous advantages. However, unrestricted autonomy would be a dangerous prospect, and meaningful human judgment must be preserved. Warfighters need risk-based, bounded autonomy, with human authority increasing as consequences intensify. In situations where the stakes are highest, the human decisionmaker must remain the final safeguard against unintended harm and irreversible outcomes. ★

Endnotes

- 1 U.S. Department of Defense, *Military and Security Developments Involving the People's Republic of China 2025*, Annual Report to Congress (Washington, DC: Department of Defense, 2025), ch. 5, "Emerging Technological Developments."
- 2 DOD, *Military and Security Developments Involving the People's Republic of China 2025*, p. 57.
- 3 People's Republic of China, Ministry of Foreign Affairs, "Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence (AI)," December 14, 2021.
- 4 DOD, *Military and Security Developments Involving the People's Republic of China 2025*, p. 57.
- 5 Danny Tobey et al., "China Releases AI Safety Governance Framework," DLA Piper, September 12, 2024; and Cole McFaul, Sam Bresnick, and Daniel Chou, *Pulling Back the Curtain on China's Military-Civil Fusion: How the PLA Mobilizes Civilian AI for Strategic Advantage* (Washington, DC: The Center for Security and Emerging Technology, Georgetown University, September 2025).
- 6 Russian Federation, "Strategy for the Development of Unmanned Aviation"; Kateryna Bondar, *How Russia Is Building a Sovereign Drone Ecosystem for AI-Driven Autonomy* (Washington, DC: Center for Strategic and International Studies, April 13, 2026); and Ioannis Sidiropoulos, "The Current State or Russian Military AI," *Strategy International*, September 22, 2026.
- 7 Bondar, *How Russia Is Building a Sovereign Drone Ecosystem for AI-Driven Autonomy*.
- 8 Steve Balestrieri, "Russia's Military Is a Drone 'Superpower,'" *National Security Journal*, June 5, 2025; and Bondar, *How Russia Is Building a Sovereign Drone Ecosystem for AI-Driven Autonomy*.
- 9 Samuel Bendett, *The Role of AI in Russia's Confrontation with the West* (Washington, DC: Center for a New American Security, May 3, 2024); and Bondar, *How Russia Is Building a Sovereign Drone Ecosystem for AI-Driven Autonomy*.
- 10 Sidiropoulos, "The Current State or Russian Military AI."
- 11 "Russian Federation" in "Lethal Autonomous Weapons Systems: Report of the Secretary-General, UN General Assembly, A/79/88," July 2024, pp. 94–97.
- 12 Bondar, *How Russia Is Building a Sovereign Drone Ecosystem for AI-Driven Autonomy*; and Russian Federation, "Strategy for the Development of Unmanned Aviation."
- 13 Michael Barnhart, "North Korea's Integration of AI Across Cyber, Economic, and Military Domains," 38 *North*, Stimson Center, February 27, 2026.
- 14 Jack Kim, "North Korea tests AI-guided missiles and artillery rockets designed for modern warfare, KCNA says," *Reuters*, May 26, 2026.
- 15 Jack Kim, "North Korea Leader Kim Jong Un Touts AI Suicide Drones, Early-Warning Aircraft," *Reuters*, March 27, 2025.
- 16 Gintaras Radauskas, "Report: North Korea ready to use AI in military operations," *Cybernews*, updated November 28, 2025.
- 17 Microsoft Threat Intelligence, "Staying Ahead of Threat Actors in the Age of AI," February 14, 2024.
- 18 The White House, "Statement by NSC Spokesperson Adrienne Watson on the U.S.-PRC Talks on AI Risk and Safety," May 15, 2024; and Kyle Chan et al., "AI Risks from Non-State Actors," Brookings Institution, January 29, 2026.

About The Mitchell Institute

The Mitchell Institute educates the general public about aerospace power's contribution to America's global interests, informs policy and budget deliberations, and cultivates the next generation of thought leaders to exploit the advantages of operating in air, space, and cyberspace.

Forum Submissions and Downloads

For more information about submitting papers or ideas to the Forum, or for media inquiries, email our publications team at forum.mitchellaerospacepower@afa.org

Copies of Forum papers can be downloaded under the publications tab on the Mitchell Institute website at <https://www.mitchellaerospacepower.org>

About the Author

Maj Gen Thomas “Tom” Taverney, USAF (Ret.), is a distinguished military leader, aerospace executive, and a leading authority on space warfare. Over a 38-year career in the United States Air Force, Gen Taverney served in meaningful launch, satellite operations, mission operations, engineering, and acquisition roles, culminating as the Vice Commander of Air Force Space Command (AFSPC).

His career reflects the evolution of space into an active operational domain, driven by his early contributions to tactical space defense. In 1969, his work on miniature homing vehicle guidance earned him the Air Force Scientific Achievement Award, which laid the groundwork for him to design the Air Force's very first Airborne Anti-Satellite (ASAT) system the following year. This tactical foresight eventually led to his formal recognition as an official Space Pioneer and his induction into both the Space Operations Hall of Fame and Space Systems Command's Schriever Wall of Honor. Throughout his active-duty service—for which he was awarded the Distinguished Service Medal—and his subsequent leadership roles in the private sector, General Taverney has remained at the forefront of orbital operations, launch, and military Space strategy.

His contributions in design and development of leading-edge technology EO/IR payloads, rapid development of systems, orbital resilience and proliferations and mission focused integration have been marked by peer honors, including the NASA Stellar Achievement and Aviation Week Awards. He has been honored with the Gen Bernard Schriever Lifetime Achievement Award and the Gen Thomas D. White Space Award. Gen Taverney continues to provide significant contributions to the Space community as an industry EO/IR payload developer, strategic advisor, author, and regular commentator on orbital security and great power competition.

