OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Governing the Algorithm: How AI Transforms National Security in a Multipolar World

Samuel Zayas
srzayas4@gmail.com

ABSTRACT

This paper argues that artificial intelligence (AI) is transforming national security not primarily through greater destructive capacity, but through three interlocking mechanisms: the accelerated tempo of decision-making, the opacity of model reasoning, and the autonomy that compresses the space for human judgment. These properties systematically undermine strategic stability by increasing misclassification risk, reducing time for interpretation and restraint, widening legal and ethical responsibility gaps, and pushing lethal workflows toward machine-speed operation. To substantiate this theoretical framework, the paper employs a structured comparative case study analysis, examining the integration of layered automation in contemporary conflicts in Ukraine and Gaza. This empirical analysis demonstrates how the compression of the "sensor-to-shooter" loop functionally narrows human validation windows in practice. The logic of compression and opacity is then extended through a focused conceptual analysis to the domain of nuclear command-and-control (NC3), where the paper contends that even advisory or decision-support automation could destabilize crisis signaling and raise the risk of inadvertent escalation. Furthermore, the paper evaluates how private-sector dominance over frontier compute and model access reshapes state sovereignty, while multipolar divergence in semiconductor supply chains enables the development of parallel, incompatible AI ecosystems—especially among nonaligned states—thus challenging traditional assumptions of U.S.-led institutional rule-setting. In response to these interconnected risks, the paper concludes by synthesizing a policy architecture centered on enforceable, testable technical and institutional guardrails. These include: an aviation-style military AI incident reporting regime; mandatory, contractually embedded TEVV (test, evaluation, verification, validation) processes; cryptographically enforceable specifications for meaningful human control; NC3 systems with "default-to-delay" logic under anomaly; calibrated export controls paired with transparency measures; deepfake crisis-preparedness protocols; and the proactive inclusion of Global South actors in technical standard-setting. Ultimately, this analysis contends that effective governance in the AI-military domain will be determined first by technical engineering defaults—such as logging schemas, evaluation playbooks, and interoperability baselines—not by treaty language alone. Preserving human judgment in an age of machine-speed warfare requires consciously designing socio-technical systems that move fast only when evidence is strong and verifiable, and that are architected to slow down automatically when uncertainty spikes (Stanley-Lockman, 2021; Osoba & Welser, 2017; Ppakter, 2024).

January 2026
Vol 3. No 1.

INTRODUCTION

Artificial intelligence (AI) has transitioned from a speculative research frontier to a constitutive technology of twenty-first-century statecraft and conflict. It is reshaping how individuals work, communicate, and perceive the world, and with even greater consequence, it is fundamentally altering how states conceive of, prepare for, and wage war. The migration of AI from commercial laboratories and digital platforms to the core of national security planning is now a defining feature of great-power competition. Applications span the full spectrum of military operations: from intelligence, surveillance, and reconnaissance (ISR) and electronic warfare (EW) to cyber defense, logistics optimization, and precision targeting. The United States and the People's Republic of China stand as the most visible and resourced competitors in this domain, each making monumental investments in frontier models, sovereign semiconductor capacity, and autonomous platforms designed to reduce dependency on continuous, slow-loop human input. U.S. export controls on advanced computing chips aim to raise rivals' costs and preserve a technological edge, but they also have the perverse effect of accelerating parallel development efforts abroad (National Telecommunications and Information Administration, 2024; Bo et al., 2022). Meanwhile, U.S. allies and partners are actively experimenting with manned-unmanned teaming (MUM-T), distributed autonomous swarms, and interconnected "kill-web" architectures that seek to compress the time from target detection to kinetic effect to an unprecedented degree (DARPA, 2020; U.S. Air Force, 2025; Department of Defense, 2022).

This rapid integration prompts a critical question: How, precisely, does military-grade AI transform the essential character of national security and strategic stability? This paper contends that the answer lies not merely in additive capability—doing old things with marginally higher efficiency—but in a qualitative transformation driven by a triad of systemic properties: acceleration, opacity, and autonomy. Together, these properties risk eroding meaningful human control, straining crisis stability, and complicating legal accountability. As AI-enabled systems become faster, more interconnected, and more autonomous, they consistently outpace the development of corresponding regulatory frameworks and meaningful human oversight mechanisms. Speed compresses decision windows to milliseconds, leaving political leaders and frontline operators with less time for deliberation, context-checking, and restraint. The sheer scale and distribution of AI systems multiply complex interactions across a mesh of sensors, fusion models, and command-and-control (C2) nodes, creating novel failure modes. Most critically, the functional opacity of many advanced models—often termed the "black box" problem—leaves operators and strategic leaders unsure of how a specific recommendation was generated or how a system will behave under novel, stressful, or adversarial conditions. This epistemic gap, widened by the relentless competitive pressure to deploy first, creates fertile ground for catastrophic misclassification, rapid unintended escalation, and a diffusion of responsibility that existing legal and ethical frameworks are ill-equipped to manage (Stanley-Lockman, 2021; Bajema et al., 2022; Osoba & Welser, 2017).

This paper argues that AI transforms national security by (1) converting time into a directly contested domain where tempo dominance confers strategic initiative; (2) widening responsibility gaps that make lawful accountability and ethical judgment harder to enforce at machine speed; and (3) shifting power and sovereignty among states, private firms, and non-state actors within an emerging multipolar, "multistack" global ecosystem. The evidence for these claims is drawn from a systematic analysis of battlefield practice in Ukraine and Gaza; from doctrinal shifts evident in U.S. and allied programs for machine-speed C2 like CJADC2; from the geopolitical dynamics of semiconductor industrial strategies and export controls; and from early—though still fragmented—international governance efforts. The goal of this analysis is not alarmism but mechanistic clarity: to trace the specific pathways through which technical performance parameters become sources of strategic risk, and to specify actionable, technically-grounded guardrails designed to preserve human judgment and strategic stability where they matter most (Stanley-Lockman, 2021; Bajema et al., 2022; Crootof, 2022).

The paper advances three core theses. First, autonomy's most consequential strategic effect is temporal: the actor who can reliably operate inside an adversary's observation-orientation-decision-action (OODA) cycle—without sacrificing system reliability or inviting catastrophic error—gains a potentially decisive advantage. Second, real accountability in the age of AI is a design and engineering property, not merely a post-hoc legal or rhetorical assertion; immutable audit logs, rigorous test-evaluation-verification-validation (TEVV) protocols, and structured incident reporting regimes must be co-designed and co-produced alongside warfighting capabilities. Third, in a world where foundational technical stacks (hardware, software, models) are diverging, effective governance travels through shared engineering artifacts—such as standardized logging schemas, evaluation playbooks, and interoperability baselines—at least as much as through diplomatic treaties or declaratory policy. This analytic frame helps reconcile and advance contemporary debates about great-power competition, corporate influence in defense, and the agency of Global South states in shaping the future of military AI (Bruun & Bo, 2025; Int'l Law Studies, 2021; Spektor, 2025).

METHODOLOGY AND SCOPE

This paper is structured as a conceptual-theoretical analysis grounded in comparative case study evidence. Its primary methodology is the structured examination of causal mechanisms—specifically, compression, opacity, and autonomy—across distinct empirical domains to build a coherent, testable argument about AI's transformational impact. The case selection of Ukraine and Gaza is deliberate and justified: these are two of the most data-rich, contemporary conflicts where the integration of commercial and military AI technologies is occurring in real-time, under intense operational pressure, and is subject to significant open-source investigative reporting. They offer contrasting but complementary vistas: a large-scale interstate conflict (Ukraine) and a dense, asymmetric urban warfare environment (Gaza). Analyzing both allows for a more robust examination of how the core mechanisms manifest across different conflict typologies. The extension of the argument to nuclear command, control, and communications (NC3) constitutes a plausibility probe, using logical inference from the established mechanisms to explore risks in a domain where empirical data is scarce but potential consequences are existential. The analysis of semiconductor supply chains and private-sector power draws on policy documents, industry reports, and economic statecraft literature to illustrate the structural constraints and enabling conditions for the deployment of military AI.

The scope of the paper is necessarily bounded. It focuses on the *implications* of AI integration for strategic stability and governance, rather than providing a comprehensive technical primer on the AI systems themselves. It acknowledges an overreliance on open-source reporting (OSINT), which, while a necessary resource for studying contemporary conflict, introduces limitations regarding verification and completeness. Some future-oriented claims, particularly regarding AI and NC3, are inherently speculative but are grounded in extrapolations from current technological trends and well-established theories of crisis stability. The findings from the selected cases are not automatically generalizable to all forms of warfare but are indicative of powerful trajectories shaping high-intensity and asymmetric conflicts. Finally, while the paper engages with International Humanitarian Law (IHL) and ethical debates, its primary focus remains on the strategic and policy dimensions of AI integration (Newton, 2025; Pusztaszeri & Harding, 2025; Wiese & Langer, 2024; Crootof, 2022).

PAPER STRUCTURE

Section 1 traces the historical and doctrinal rise of autonomous and semi-autonomous systems in security contexts, establishing why autonomy's primary strategic value lies in tempo control. Section 2 analyzes the transformation of command and control (C2) into machine-speed decision cycles, repositioning the human operator as an exception handler and raising critical human-factors challenges. Section 3 examines the structural drivers of risk inherent to military AI: model opacity, distribution shift, and the presence of an intelligent adversary capable of adversarial attacks. Section 4 presents the core comparative case studies of Ukraine and Gaza, detailing how layered automations are compressing human validation windows in live combat. Section 5 extends the logical framework to nuclear command, control, and communications (NC3), arguing for the necessity of "default-to-delay" architectures and dual-path adjudication to preserve crisis stability. Section 6 reconceptualizes human judgment as a depletable cognitive resource in machine-tempo environments and specifies design features—graceful degradation, escalation gates, uncertainty-aware user interfaces—required to protect it. Section 7 synthesizes a holistic risk-mitigation framework, covering accountability-by-design, pragmatic oversight via technical standards, responses to key counter-arguments, TEVV as a contractual obligation, NC3 safeguards, and deepfake preparedness.

Section 8 analyzes the problem of private corporate power in public defense and proposes remedies to avoid "sovereignty by subscription." Section 9 details the geopolitics of semiconductors and supply chains, demonstrating how export controls catalyze multistack workarounds and parallel ecosystem development. Section 10 situates these technical and economic dynamics within broader debates over global order, arguing for a shift in focus from treaty-based institutions to "artifact governance." Section 11 operationalizes the principle of "meaningful human control" into a set of falsifiable, testable technical specifications. Section 12 outlines the institutional "safety stack" required to turn principles into practice: incident regimes, adversarial red-teaming, and independent test ranges. The paper concludes with a targeted, integrated policy framework (Section 13) and a final summary of findings.

Because autonomy only reshapes strategic outcomes when it alters tempo within decision architectures, we begin by charting the entry of autonomous AI into security systems and why this shift makes the transformation of command and control the central problem of modern warfare.

THE RISE OF AUTONOMOUS AI IN SECURITY SYSTEMS

The integration of artificial intelligence into the battlespace represents a paradigmatic shift in the character of military operations, not through the wholesale replacement of human warfighters, but through a fundamental reallocation of cognitive and temporal labor—changing where, when, and how human judgment matters most. Traditional automated systems execute explicit, pre-programmed instructions coded by engineers. In contrast, modern AI systems, particularly machine learning (ML) models, *infer* patterns and rules from vast datasets, generalize—often imperfectly—to novel contexts, and can act with varying degrees of operational independence across functions including ISR, EW, cyber defense, logistics, and targeting (Stanley-Lockman, 2021; National Telecommunications and Information Administration, 2024). This shift from deterministic automation to probabilistic autonomy is profound. Furthermore, because the vast majority of foundational AI research and development originates in the commercial technology sector, its migration into the defense ecosystem brings with it a unique set of socio-political baggage: Silicon Valley developer norms and ethics, consumer brand reputations, friction from export controls on dual-use technologies, and volatile public legitimacy. This social and political context often matters as much as the technical specifications of the hardware and software.

The U.S. Department of Defense’s Project Maven (2017) stands as the canonical example of this complex transition. Technically, the program successfully paired commercial-grade computer vision algorithms with feeds from military drones to triage massive volumes of full-motion video (FMV), sharply reducing the time human analysts needed to screen footage and significantly increasing intelligence throughput. Yet, the public and political reaction to Maven taught an equally potent lesson about the political economy of military AI. Thousands of Google employees protested the company’s involvement, citing ethical concerns over the use of AI for warfare, and the internal revolt ultimately contributed to the company’s decision not to renew the contract. Project Maven thus illustrates a dual reality: (1) AI can deliver substantial operational value by automating perceptual tasks and accelerating the cueing of human attention; and (2) the commercial origins of AI—with attendant workforce agency, consumer brand sensitivities, and shareholder pressures—can constrain military adoption as effectively as a failed technical test (Risk Innovation Nexus, 2019).

The conceptual and operational line between assistive decision-support tools and genuinely autonomous actors is blurring rapidly in actual field use. Reporting on the Turkish Kargu-2 loitering munition’s alleged use in Libya in 2020 suggested the system may have operated in a "fire, forget, and find" mode, engaging targets without requiring continuous human control, even if a human had authorized the mission at its outset (Nasu, 2021). In the ongoing war in Ukraine, both Ukrainian and Russian forces have innovated a form of distributed, algorithmically accelerated warfare. They blend first-person-view (FPV) drones, real-time object-recognition models running on mobile phones or edge devices, and loitering munitions into integrated kill chains. The process—where sensors discover potential targets, AI models classify them, mobile apps aggregate and geolocate sightings, and artillery or drone operators act—occurs in cycles compressed from hours or minutes to seconds. This compression is conditioned by the constant electronic warfare duel, where each side jams and spoofs the other's sensors and communications. The most significant innovation here is often cumulative rather than singular: numerous small, locally rational automations, when stacked together, create a system-of-systems that *functionally* narrows the time and cognitive space available for meaningful human intervention, even if no single link is officially designated as "fully autonomous" (Newton, 2025; Pusztaszeri & Harding, 2025).

The domain of air-to-air combat presents perhaps the starkest vision of this human-machine pivot. DARPA's AlphaDogfight Trials in 2020 demonstrated that an AI agent could defeat an experienced human F-16 pilot in a simulated dogfight. The AI's victory stemmed not from superior creativity or "thinking," but from its ability to calculate, predict, and commit to maneuvers at speeds far exceeding human neural and physiological reaction times. Building on this, the U.S. Air Force's Collaborative Combat Aircraft (CCA) program aims to field AI-enabled unmanned "wingmen" that team with piloted fighter jets. These autonomous aircraft would scout ahead, conduct electronic jamming, and potentially strike targets under broad human direction. The strategic value is not merely force multiplication; it is tempo multiplication. When an AI wingman can generate and propose viable tactical options in milliseconds, the human pilot's role fundamentally shifts from being the primary creator of courses of action to being a selector and validator among machine-generated options. This shift represents the core of AI's strategic transformation: it creates executable choices that humans would never have had the time to conceive or evaluate under the pressures of combat (DARPA, 2020; U.S. Air Force, 2025).

In this emerging AI-human strategic paradigm, time itself becomes a contested domain. In previous eras of warfare, decisive advantage accrued from factors like industrial production capacity, geographic positioning, or intelligence asymmetries. In the emerging era, a central source of advantage is accruing from temporal dominance—the capacity to consistently operate inside an adversary's observation-orientation-decision-action (OODA) cycle while maintaining a high degree of system reliability. This is why autonomy matters strategically: not because it is technologically flashy, but because it transforms time from a neutral backdrop into a scarce, maneuverable, and weaponizable resource. This realization also explains this paper's analytical structure: to understand why and when autonomy becomes strategically destabilizing, we must first study the command and control architectures where tempo is translated into concrete action. We therefore turn next to an examination of how C2 is being re-engineered for machine-speed decision cycles.

COMMAND, CONTROL, AND MACHINE-SPEED DECISION CYCLES

The evolution of modern command and control (C2) under the influence of AI represents more than a software upgrade; it signifies a conceptual shift from episodic, deliberative staff work to a state of continuous, algorithmic triage of the battlespace. The U.S. Department of Defense's vision for Combined Joint All-Domain Command and Control (CJADC2) encapsulates this ambition. It aims to create a seamlessly interconnected "sensing grid" across all domains (space, air, land, sea, and cyberspace), where AI-powered data fusion engines synthesize information and decision aids "nominate" optimal actions faster than an adversary can perceive or react. Proponents argue that this acceleration can save lives and win engagements by intercepting inbound threats earlier and coordinating the effects of geographically dispersed platforms under fire with unprecedented speed and precision. Critics, however, warn that this acceleration inherently raises the risk of false positives, compresses or eliminates crucial deliberation time, and may reduce human authorization to a perfunctory "rubber stamp" on decisions that have already been algorithmically framed and narrowed (Bajema et al., 2022; Department of Defense, 2022).

Programs like the U.S. Air Force's Collaborative Combat Aircraft (CCA) and its predecessor Skyborg are tangible steps toward this vision, developing the hardware and autonomy kernels for AI wingmen. At sea, DARPA's Sea Hunter unmanned surface vessel and the Navy's Orca Extra Large Unmanned Undersea Vehicle (XLUUV) prototype represent efforts to inject persistent, low-cost autonomy into maritime domains, particularly for tasks like submarine tracking. Each of these autonomous or semi-autonomous nodes adds new capability and resilience to what strategists term the "kill web." However, each new node also introduces fresh potential points of failure and novel interaction risks. A mis-specified rule in a targeting recommender, a computer vision classifier fooled by adversarial "noise," or a microseconds-level clock synchronization glitch can now propagate decisions and kinetic effects at machine speed across the network (U.S. Air Force, 2025; U.S. Navy, 2025; U.S. Government Accountability Office, 2022).

Within these machine-tempo environments, the human operator's cognitive burden does not disappear; it mutates. The human is increasingly repositioned as an exception handler rather than the primary decision engine. Two critical human-factors failure modes loom large in this context: (1) Automation Bias, where an operator over-relies on a machine's output because its display appears precise or carries a high numerical "confidence" score, and (2) Alert Fatigue, where a constant stream of machine-generated alerts—many of which are false positives—conditions the operator to dismiss warnings reflexively. Both failure modes systematically undermine the operator's interrogative stance—the crucial capacity to pause and ask, "What else could explain this data? What evidence would change my mind?" Therefore, the design of user interfaces (UX) for AI-enabled C2 becomes a strategic imperative. Effective interfaces must expose not just a single recommendation, but salient features of the data, plausible alternative interpretations, the lineage of sensor and model inputs, and a decomposition of the system's uncertainty—rather than presenting a single, potentially misleading scalar confidence value. Without these cognitive affordances, even a well-calibrated AI model becomes a strategic liability because the human in the loop lacks the time and tools to interrogate its outputs meaningfully (Osoba & Welser, 2017; Ppakter, 2024).

This compression of decision loops also changes the points at which governance and human oversight can be most effective. If human review occurs only at the tail end of a processing chain where options have been pre-selected and the choice has been narrowly framed by the algorithm, then the legal requirement for "authorization" risks becoming a ceremonial step. To preserve meaningful control, friction and deliberation must be intentionally designed upstream in the process. This can be achieved through engineered gating functions that automatically require additional sensor corroboration, senior-officer review, or simply slow the system's tempo when key uncertainty thresholds are crossed (e.g., low confidence, sensor disagreement, or entry into geofenced restricted zones). In practice, this means building safety policy directly into software pipelines: a responsible AI-enabled system should be designed to *slow itself down* when the operational picture becomes ambiguous or contested (Osoba & Welser, 2017; Crootof, 2022).

Before examining how these dynamics play out on actual battlefields, it is essential to understand the inherent technical vulnerabilities of AI systems that make acceleration so perilous. The next section examines the structural drivers of risk: opacity, distribution shift, and the intelligent adversary.

THE DANGER OF THE UNKNOWN: OPACITY, DISTRIBUTION SHIFT, AND ADVERSARIES

The performance of AI systems, particularly deep learning models, is often accompanied by profound epistemic challenges for their human operators. Many state-of-the-art models are functionally opaque: even their engineers cannot reliably trace or explain why a specific input yielded a specific output. In commercial applications like product recommendations, this may be a tolerable trade-off for performance. In national security, where decisions can lead to lethal outcomes and escalation, opacity acts as a powerful instability multiplier. Large language models (LLMs) are known to "hallucinate" plausible but false information; computer vision models can encode and amplify societal biases related to demographics or environmental contexts; multi-modal systems can compose errors across different data types (e.g., image and text). Crucially, performance measured on clean, curated civilian benchmarks provides almost no reliable predictor of how a system will perform under the chaotic, non-stationary conditions of the battlefield (Osoba & Welser, 2017; Trendsresearch.org, 2025).

Distribution shift—the divergence between the data a model was trained on and the data it encounters in real-world deployment—is the norm, not the exception, in warfare. Combat creates a uniquely non-stationary environment: changing weather, smoke and dust, visual clutter, novel camouflage techniques, physically degraded sensors, emergent adversary tactics, and intense electronic warfare interference. A model trained to recognize tanks in clear California desert imagery may fail utterly to identify the same object in Ukrainian mud, under camouflage netting, or amidst the rubble of Gaza. Accuracy can collapse precipitously and without warning (Bajema et al., 2022; Osoba & Welser, 2017).

A thinking adversary will actively seek to exploit these vulnerabilities through adversarial machine learning. This includes adversarial perturbations—subtle, often human-imperceptible alterations to an input (e.g., a pixel pattern on a vehicle) designed to cause a classifier to mislabel it—and data poisoning attacks, where an adversary corrupts the training data to cause a system to fail in specific, latent ways. These attacks are often orders of magnitude cheaper than kinetic countermeasures yet can trigger disproportionately large effects, especially if they feed into automated "sensor-to-shooter" loops that act on misclassifications before a human can intervene (Bajema et al., 2022; Osoba & Welser, 2017).

The most dangerous systemic failures are rarely the product of a single model's error. They emerge from complex interactions and feedback loops within a system-of-systems. A misclassification by a perception model feeds an incorrect premise to a planning model, which then proposes a flawed course of action to a C2 recommender. Small errors can multiply as they cascade through these interconnected loops. Therefore, safety engineering for military AI must focus on closed-loop behavior in instrumented, realistic test environments, not merely on optimizing single-model accuracy on static datasets (Osoba & Welser, 2017).

Adding another layer of complexity, information operations and synthetic media act as coupled cognitive stressors. A strategically released deepfake—a convincing AI-generated video or audio recording—at a moment of crisis can delay authorization processes, trigger premature defensive actions, or sow paralyzing doubt within a chain of command. When such synthetic signals coincide with ambiguous sensor data or disruptive cyber incidents, the lag in attribution and verification directly translates into escalation risk (Twomey et al., 2023).

These are not merely theoretical concerns drawn from laboratory experiments. In live conflicts, the integration of automation has moved from demonstration videos to integral parts of battle rhythm. The following section presents a structured comparative analysis of the conflicts in Ukraine and Gaza to examine how the mechanisms of compression and opacity manifest in practice, revealing the points where the concept of "human control" is being stretched to its practical limits.

BATTLEFIELD PRACTICE: UKRAINE, GAZA, AND THE EDGES OF HUMAN CONTROL

Case studies serve to anchor theoretical mechanisms in observable reality. The conflicts in Ukraine (2022-present) and Gaza (2023-2024) provide critical, real-world laboratories for examining how ostensibly "reasonable" micro-automations combine into integrated workflows that progressively shrink the windows for human validation and judgment.

THE CASE OF UKRAINE: ALGORITHMICALLY ACCELERATED ATTRITION

Since the full-scale invasion in 2022, the war in Ukraine has evolved into a testing ground for decentralized, AI-enabled warfare. Both sides have implemented layered automation across the kill chain. At the tactical edge, lightweight object-detection models run on smartphones or single-board computers attached to small drones (FPV and reconnaissance), automatically classifying potential targets (vehicles, personnel, equipment). These detections are fed into mobile applications (like Ukraine's use of Delta or Kropyva) that aggregate sightings from multiple drones and operators, geolocate targets on digital maps, and generate prioritized target lists for artillery units or drone strike teams. Electronic warfare systems add another automated layer, constantly scanning and jamming enemy communications and navigation signals, while also adapting to counter enemy EW. No single link in this chain is officially "fully autonomous"; a human technically authorizes each strike. However, the decision latency from detection to firing has collapsed from tens of minutes to mere seconds in many engagements. As one analyst noted, the human is "in the loop," but the loop has become so fast and cognitively demanding that the human's role is reduced to rapid verification under extreme time pressure, often trusting the algorithmic cueing (Newton, 2025; Pusztaszeri & Harding, 2025).

A poignant example of the information warfare dimension was the deepfake video of President Volodymyr Zelensky that circulated briefly in March 2022, appearing to call on Ukrainian soldiers to lay down their arms. It was quickly debunked, but it served as a stark preview of how synthetic media could be weaponized to create confusion, delay decisions, and undermine command integrity at a critical moment. As generative AI quality improves, the time window for effective debunking will narrow, forcing militaries to operate in an environment where any piece of information, including apparent orders from commanders, must be treated as potentially synthetic (Twomey et al., 2023).

THE CASE OF GAZA: SCALE, THROUGHPUT, AND THE "LAVENDER" SYSTEM

Reporting by investigative journalists on the Israeli Defense Forces' (IDF) use of an AI-based targeting system reportedly nicknamed "Lavender" during the 2023-2024 conflict in Gaza sparked intense international debate about scale, accuracy, and the nature of human review. According to these reports, the system analyzed masses of surveillance data to generate "kill lists" of tens of thousands of individuals alleged to be low-level Hamas operatives, assigning them a statistical score. Reports further suggested that for certain categories of targets, human review was exceptionally brief, sometimes lasting only seconds per target, effectively serving as a "rubber stamp" (Wiese & Langer, 2024; Crootof, 2022).

Without verifying the specific, contested details of these reports, the "Lavender" case illustrates a structural risk inherent to AI-enabled targeting at scale. When algorithms can generate potential targets orders of magnitude faster than human analysts can meaningfully vet them, it fundamentally restructures the workflow and the psychology of decision-making. Even in systems where humans retain formal legal authority, they face immense throughput imperatives ("clear the queue") and are presented with interfaces that imply algorithmic precision and certainty. In such an environment, human review can devolve into a perfunctory box-checking exercise. When this happens, the practical locus of control has already shifted toward the machine, regardless of legal or doctrinal statements to the contrary (Crootof, 2022; Ppakter, 2024).

The broader lesson from both cases, echoing the earlier use of drones and loitering munitions by Azerbaijan in the 2020 Nagorno-Karabakh war, is that partial autonomy, when tightly integrated with ISR and C2, transforms operational tempo. The system does not need to be perfect or fully autonomous; it only needs to be "good enough" to process information and cue actions at a pace that overwhelms an adversary's decision cycle and, potentially, the ethical and legal review processes of the user (National Telecommunications and Information Administration, 2024).

If compressed decision cycles introduce profound risks in conventional warfare, their introduction into the nuclear realm multiplies the potential consequences by orders of magnitude. We therefore extend the analytical framework to the sanctum of strategic stability: nuclear command, control, and communications.

NUCLEAR COMMAND, CRISIS STABILITY, AND THE AUTOMATION TEMPTATION

Nuclear command, control, and communications (NC3) systems are already designed to operate under extreme time pressure and pervasive ambiguity, balancing the need for survivability and reliable retaliation with robust safeguards against accidental or unauthorized launch. Introducing AI and machine learning as decision *aids* into this environment—whether for early warning, threat assessment, or decision-support—without meticulously engineered safeguards poses grave risks to crisis stability. The core mechanisms of compression and opacity could have existential consequences in this domain.

For instance, AI-enabled data fusion engines, designed to synthesize inputs from satellites, radars, and other sensors, might output a single, clean composite track with a high-confidence value. This presentation could mask underlying uncertainty or dissent among individual sensors. A national leader under the extreme stress of a potential nuclear crisis could dangerously overweight such a seemingly definitive display. Unless the AI interface is explicitly designed to surface uncertainty decomposition—showing which sensors agree, which disagree, and how confidence has evolved over time—it invites catastrophic automation bias. The political declaration that "a human must always authorize nuclear use" is a necessary but insufficient safeguard if the opaque tools that shape that human's perception of reality systematically bias them toward a single, potentially false, conclusion (Osoba & Welser, 2017; Federation of American Scientists, 2025; Ppakter, 2024).

Furthermore, the mere perception that an adversary is incorporating AI for faster decision-making could generate dangerous preemptive pressures. If one side believes its command system is slower, it may feel compelled to delegate more authority to machines or to adopt risky "launch-on-warning" postures to avoid being disarmed. This dynamic, known as the "flash war" risk, echoes classic arms-race instability but at machine speeds (Bajema et al., 2022; Federation of American Scientists, 2025; Yeung et al., 2021).

Therefore, the design philosophy for any AI touching NC3 must invert the dominant logic of acceleration. NC3 aids should be engineered with "default-to-delay" or "slow-down" protocols under conditions of anomaly. This means systems should be designed to: automatically elevate dissenting sensor views for human attention; require dual-path confirmation from independent data sources for high-confidence alerts; and trigger mandatory senior-officer review when specific confidence or sensor-consistency thresholds are breached. Explicit technical and doctrinal prohibitions on any autonomous *escalation* pathways (e.g., a system that can recommend moving to a higher alert status on its own) are a minimum safeguard. The overriding principle must be that AI in NC3 should expand decision time and clarify ambiguity, not compress and obscure it (Stanley-Lockman, 2021; Federation of American Scientists, 2025; Yeung et al., 2021; Ppakter, 2024).

The NC3 discussion crystallizes the paper's central ethical and strategic concern: the preservation of human judgment under machine tempo. We now analyze judgment not as a static principle, but as a depletable cognitive resource that systems must be designed to protect.

THE LOSS OF HUMAN JUDGMENT

War has historically been, in part, a conversation between adversaries—a series of signals, pauses, feints, and gestures that allow for bargaining, de-escalation, and the interpretation of intent. During the Cuban Missile Crisis, deliberate slowing of communications and decision processes by both President Kennedy and Chairman Khrushchev created the essential space for negotiation that averted nuclear war.

Machine-paced recommenders, optimized for statistical "optimality" in strike packages, inherently tend to compress this space for signaling and reflection. They privilege rapid, decisive action over deliberate pause.

Operators confronting finite time, complex displays, and a stream of machine-ranked options are at risk of two pathologies: automation bias and alert fatigue. Both undermine the human's ability to interrogate, to re-frame, and to resist tempo itself. The problem is not the human's presence but the human's cognitive absorption threshold. If the system produces more "decisions" than the operator can comprehend and contest within the time allotted, then the human is technically "in the loop" while being effectively outpaced (Osoba & Welser, 2017; Crootof, 2022; Ppakter, 2024).

The remedy is to embed features that protect judgment. This includes graceful degradation, where systems automatically step down from autonomy to supervision or direct control when uncertainty spikes; escalation gates that require senior review for high-consequence actions; and explainable user interfaces that present salient features, counterfactual alternatives, and clear uncertainty decomposition rather than a single confidence score. The key metric is simple to state but hard to achieve: can trained operators consistently understand, challenge, and override the system within the operational time frame? If a system cannot meet this standard, it fails the test of meaningful human control, even if its technical performance on benchmarks is stellar (Osoba & Welser, 2017; Crootof, 2022; Ppakter, 2024).

Preserving judgment and managing risk requires enforceable accountability and workable governance. We therefore turn to mitigation frameworks: how to make responsibility traceable, rules scalable, and safety a condition of procurement.

MITIGATING RISK

Risk mitigation in AI-enabled defense is not "add more signatures" or "hold more reviews." It is a structural re-architecture of time, evidence, and authority. Because loops have shrunk, governance must move from paper to pipeline.

ACCOUNTABILITY AND REGULATION

Responsibility cannot be asserted after the fact; it must be instrumented at the time of action. That means immutable logs (inputs, outputs, timestamps, version hashes), provenance trails (sensor and model lineage), and model cards bound to deployments. Without traceability, the Law of Armed Conflict (LOAC) becomes non-enforceable at machine speed. Procurement must make such artifacts deliverables, not "nice-to-have" appendices (Osoba & Welser, 2017; Crootof, 2022).

Contracts should define developer, integrator, and operator obligations with auditable handoffs: who owns data quality, adversarial testing, override thresholds, and incident reporting. Liability should map to these partitions.

GLOBAL OVERSIGHT AND REGULATION

CCW negotiations on lethal autonomous weapon systems (LAWS) stall over definitions and verification, while the ICRC advocates prohibitions on unpredictable systems and those trained to target persons (Bruun & Bo, 2025; Sati, 2023). Given AI's dual-use diffusion, global governance may advance faster through technical standards—logging minima, evaluation schemas, uncertainty displays—than through universal treaties. Coalitions can adopt and export these standards as defaults. Instruments such as NATO's DIANA and NIF, and AUKUS Pillar II, can publish evaluation playbooks, incident taxonomies, and minimum logging schemas that partners (and even nonaligned states) can adopt or adapt. In a practice-led world, artifacts are governance (Int'l Law Studies, 2021; Simmons-Edler et al., 2025; CSET, 2024; United Nations, 2024).

COUNTER-ARGUMENTS AND LIMITATIONS

Some analysts argue that speed saves lives, claiming that faster machine-generated assessments reduce the time between detection and response. While this is partially true, acceleration only produces safer outcomes when paired with uncertainty-aware user interfaces and rigorous TEVV processes; without these safeguards, increased speed simply magnifies the consequences of error and escalation risk (Stanley-Lockman, 2021; Department of Defense, 2022; Osoba & Welser, 2017).

Another common argument is that adversaries will not self-bind, making unilateral restraint pointless. However, even if rivals pursue aggressive automation, bounded autonomy still reduces the likelihood of catastrophic failures—especially in nuclear-adjacent systems—and coalitions can create de facto norms by conditioning interoperability and technology access on the adoption of shared safety artifacts (Bruun & Bo, 2025; Ppakter, 2024; Int'l Law Studies, 2021; Simmons-Edler et al., 2025; CSET, 2024; United Nations, 2024).

A third concern is that detailed logs might expose sensitive methods or operational tradecraft. Yet this fear is overstated: tiered-access controls and cryptographic protections can safeguard classified techniques while still enabling accountability and after-action learning. The aviation sector demonstrates that incident-sharing regimes can drive safety improvements without revealing strategic vulnerabilities (Greipl et al., 2024).

Open-source data from ongoing conflicts is contested; nonetheless, that uncertainty strengthens the case for independent TEVV and incident regimes to replace anecdote with evidence (Newton, 2025; Pusztaszeri & Harding, 2025; Wiese & Langer, 2024; Crootof, 2022).

TEVV AS A CONTRACTUAL STANDARD

Make test, evaluation, verification, validation a pay-gated deliverable. Coverage should include distribution-shift trials (weather, clutter, sensor damage), adversarial and poisoning tests, closed-loop live-virtual-constructive (LVC) runs where perception, planning, and C2 interactions are exercised, and human-factors evaluations that measure explanation absorption and override latency. In alliances, TEVV evidence becomes an asset: the actor who can prove reliability gains coalition trust and diplomatic leverage (U.S. Navy, 2025; U.S. Government Accountability Office, 2022; Int'l Law Studies, 2021; Simmons-Edler et al., 2025).

NC3 SAFEGURADS

Impose dual-channel sensing and analytics, explainable diagnostic views, uncertainty decomposition, default-to-delay behaviors under anomaly, explicit prohibitions on autonomous escalation, and senior review for any automation touching warning or decision aids. Instrument dashboards to show dissent, not just consensus (Stanley-Lockman, 2021; Federation of American Scientists, 2025; Yeung et al., 2021; Ppakter, 2024).

DEEPFAKE PREPAREDNESS

Fund media forensics, require cryptographic signatures for official communications, build rapid-response debunk teams, and train leaders to operate in "mixed-signal" crises where false narratives and ambiguous telemetry collide. Include synthetic media injections in wargames so decision-makers practice corroboration under pressure (Twomey et al., 2023).

Governance interacts with market structure because private firms own much of the capability stack. We therefore analyze private power and remedies to avoid "sovereignty by subscription."

PRIVATE POWER IN PUBLIC DEFENSE

Commercial firms now build, host, and update much of the AI that militaries want: frontier models, API gateways, cloud stacks, labeling pipelines, autonomy kits. This introduces two structural risks.

If a proprietary model contributes to an unlawful outcome, liability scatters across developer, integrator, and operator. Without contractually mandated logs and evaluation artifacts, after-action attribution becomes guesswork (Crootof, 2022).

When mission workflows live behind vendor gates—quota limits, pricing knobs, model updates—governments inherit vendor priorities and workforce politics. A single policy change by a platform should not be able to alter a nation's force posture.

States should first invest in sovereign computers, funding public inference and training capacity for critical national-security uses. They should also enforce portability by requiring open evaluation harnesses, data-residency controls, and containerized deployments so that mission workflows can move across vendors without needing to be rebuilt from scratch. A third priority is traceability, mandating minimum logging schemas and strong non-repudiation so that actions are cryptographically bound to specific actors. In addition, governments need local fallbacks that allow crisis-mode inference to run even when systems are disconnected from the cloud. Finally, they should pursue diversification, expanding supplier pools through instruments such as DIANA and NIF so that no single vendor becomes a single point of strategic failure. Private power is constrained by computers and supply chains. Export controls, indigenous fabrication, and memory bottlenecks are reshaping the map. We turn next to chips and multipolar workarounds (Int'l Law Studies, 2021; Simmons-Edler et al., 2025; CSET, 2024; United Nations, 2024).

CHIPS, SUPPLY CHAINS, AND MULTIPOLAR WORKAROUNDS

Computing is the binding constraint for modern AI. Controls on top-tier GPUs increase costs and scarcity for targeted actors, but they also accelerate workarounds: local fabrication at mature nodes, precision formats tuned to available silicon, and pushes toward high-bandwidth memory. The result is stack divergence: chips, interconnects, memory, compilers, and models co-optimized into parallel ecosystems (Bo et al., 2022).

Export controls intended to slow a rival can catalyze the formation of self-reliant stacks. India's public computer investments and Brazil's investment-plus-regulation model show nonaligned strategies that avoid dependency on any single bloc. Gulf sovereign funds back vertically integrated clusters with embedded safety harnesses as a national differentiator. Power accrues not only to states, but to ecosystems—the technical and institutional stacks that others plug into (Podar & Colijn, 2025; Toutoungi, 2025).

If technical stacks now shape influence, world-order debates must account for ecosystem power. We position Spektor's and Cooley & Nexon's arguments within this terrain.

AI AND THE GLOBAL ORDER: SPEKTOR VS. COOLEY & NEXON

Multipolarity—properly harnessed—expands agency for the Global South. By bargaining across blocs, building local capacity, and asserting preferences in standards bodies, nonaligned states can secure better terms and avoid structural dependency (Spektor, 2025).

An "America First" posture that weakens institutions undercuts U.S. structural power and opens rule-making space for authoritarian actors. They are right to foreground institutions. But the analysis must widen: in AI, engineering defaults become de facto law. The actor who exports logging schemas, evaluation playbooks, and interop baselines shapes practice long before treaties mature (Int'l Law Studies, 2021; Simmons-Edler et al., 2025; CSET, 2024; United Nations, 2024; Podar & Colijn, 2025; Toutoungi, 2025; Devitt, 2021; Spektor, 2025).

Institutions still matter, but the center of gravity is shifting toward artifact governance: telemetry standards, test ranges, incident regimes, and interop APIs that codify norms in code and process. Democracies must therefore (1) sustain alliances and (2) compete in engineering arenas.

Values must be translated into specifications to shape procurement and doctrine. We therefore define meaningful human control as a testable, enforceable set of requirements.

FROM PRINCIPLES TO PRACTICE: DEFINING MEANINGFUL HUMAN CONTROL

"Meaningful human control" only matters if it is falsifiable in testing and enforceable in contracts. The following requirements make the principle operational for acquisition officers, engineers, and commanders:

1. Temporal bounds. Narrow engagement windows; automatic abort on communications degradation or verified confidence shortfall. Log timing from recommendation to authorization to effect (Bruun & Bo, 2025; Ppakter, 2024).
2. Geographic bounds. Geofence autonomous effects away from civilians and sensitive infrastructure; where feasible, confine autonomy to anti-material missions, not human targets (National Telecommunications and Information Administration, 2024; Int'l Law Studies, 2021).
3. Target-profile limits. Prohibit autonomous engagement of persons. For material, require narrowly specified, auditable profiles (e.g., RF bands, thermal envelopes) with negative lists that force abstention under ambiguity (National Telecommunications and Information Administration, 2024; Bruun & Bo, 2025).
4. Traceability and audit. Mandate immutable, mirrored logs of inputs, outputs, model versions, prompts/overrides, and timing; bind logs with cryptographic signatures across the chain of custody (Osoba & Welser, 2017; Crootof, 2022).
5. Graceful degradation. When cross-sensor divergence or uncertainty spikes, step down autonomy; if thresholds persist, lock to supervised or manual modes until reset under controlled conditions (Osoba & Welser, 2017).

6. Human-factors UX. Interfaces must explain why (salient features), what else (alternatives/counterfactuals), and how sure (uncertainty decomposition). Instrument explanation absorption (how quickly trained operators can restate rationale) and override latency.

Mechanization: token-gated autonomy. Bind autonomy modes to cryptographically signed engagement tokens issued by authorized human decision authorities. Tokens expire quickly, can be revoked mid-execution, and encode temporal/geographic envelopes. Keep physically authoritative local overrides (not just network commands) so that in degraded comms a human can still abort.

Requirements demand infrastructure. We therefore build the safety stack that turns these specs into repeatable practice.

BUILDING THE SAFETY STACK: INCIDENTS, RED-TEAMING, AND INDEPENDENT TEST RANGES

The safety stack is the institutional counterpart to capability. Its purpose is not to slow innovation, but to make fast iteration survivable. The four pillars are as follows:

1. Adversarial red-teaming at model and system-of-systems levels. Probe distribution shift (weather, clutter, damaged sensors), adversarial perturbations, and data poisoning. Include closed-loop tests where model outputs feed downstream recommenders to surface compounding errors (Osoba & Welser, 2017).
2. Independent test ranges (live-virtual-constructive; hardware-in-the-loop) with dense instrumentation to capture edge-case telemetry and build reusable regression datasets. Autonomy must be evaluated with real latencies, jitter, and EW effects (U.S. Navy, 2025; U.S. Government Accountability Office, 2022).
3. Confidential incident regime with safe-harbor reporting. Require vendors and services to file failures, surprises, and near misses, bundling standardized telemetry (logging schema, version hashes, operator trace). Share anonymized findings among coalition partners (Greipl et al., 2024).
4. Procurement incentives that pay for evidence, not just performance. Contract for coverage metrics (distribution-shift, adversarial, closed-loop), red-team artifacts, and human-factors results; include clawbacks for undisclosed failure modes (U.S. Government Accountability Office, 2022; Int'l Law Studies, 2021; Simmons-Edler et al., 2025).

Two enablers cut across the entire safety stack. The first is data governance, which requires rigorous provenance tracking, consent management, cryptographic hashing, red-team seed quarantine, and high-quality labeling oversight. These practices ensure that training and evaluation datasets remain trustworthy, auditable, and resistant to manipulation. The second enabler is people: militaries and partner institutions must embed specialized roles such as range engineers, red-teamers, human-factors evaluators, and incident analysts from the outset to ensure that safety processes are integrated directly into development and testing rather than added retroactively.

Safety must also be measurable. Over a one-year evaluation cycle, programs should assess the proportion of autonomy functions using schema-compliant logs, the total hours of range time accumulated across environments and test types, and the number of incident reports filed, triaged, and remediated. They should additionally track operator explanation-absorption and override-latency distributions, along with regression pass rates after model or system updates (U.S. Navy, 2025; U.S. Government Accountability Office, 2022; Osoba & Welser, 2017; Greipl et al., 2024). These metrics are auditor-checkable and ensure that safety is tied directly to delivery rather than treated as an afterthought.

Scaling beyond alliances. Publish a v1 logging schema (model/version hashes, input pointers, uncertainty decomposition, operator trace, token issuer), an incident taxonomy (misclassification, mis-coordination, human-factors, adversarial, comms degradation, escalation-path breach), and evaluation playbooks (scenario bundles with expected failure signatures). These artifacts enable partners—including nonaligned states—to adopt compatible safety baselines without waiting for a universal treaty (Bruun & Bo, 2025; Greipl et al., 2024; Sati, 2023; Int'l Law Studies, 2021; Simmons-Edler et al., 2025; CSET, 2024; United Nations, 2024). In a practice-led world, artifacts seed norms.

Having established mechanisms, risks, and infrastructure, we now synthesize policy: how to bound machine-speed risks while competing effectively in a multipolar, multistack world.

POLICY FRAMEWORK: PRACTICAL GUARDRAILS FOR A MULTIPOLAR AI WORLD

This framework accepts strategic competition as a fact and channels it into auditable, bounded, human-anchored pathways. It is designed for major powers, allies, and capable nonaligned states.

1. Establish an AI incident regime (aviation-style). Mandate reporting of significant failures and near misses across development, test, and operations. Standardize submissions (model cards, configuration fingerprints, immutable logs, red-team artifacts). Create allied clearinghouses for anonymous sharing under safe harbor (Greipl et al., 2024). This institutionalizes learning and creates a feedback loop from practice to policy.
2. Make TEVV non-negotiable and pay-gated. Tie payment milestones to coverage metrics: distribution-shift stress tests, adversarial and poisoning evaluations, closed-loop LVC trials, and human-factors assessments (explanation absorption, override latency). Require third-party or cross-program audits for high-consequence systems (U.S. Navy, 2025; U.S. Government Accountability Office, 2022; National Telecommunications and Information Administration, 2024; Simmons-Edler et al., 2025).
3. Codify meaningful human control in doctrine and law. Embed temporal/geographic bounding, target-profile limits (no autonomous engagement of persons), graceful degradation triggers, and traceability into service CONOPS, acquisition specifications, and Article 36 reviews; harmonize coalition minimums so the strictest partner's safeguards set the baseline (Bruun & Bo, 2025; Ppakter, 2024).

4. Align export controls with calibrated transparency. Maintain controls on high-end computers, but couple them with track-and-trace transparency for large training runs above set thresholds in military applications. Use use-case-based obligations (e.g., tighter requirements for autonomous targeting than for logistics) to avoid blanket embargoes that harden rival stacks (Bo et al., 2022).
5. Harden NC3-adjacent software against automation bias. Mandate dual-channel adjudication, explainable diagnostic views, uncertainty decomposition, and default-to-delay behaviors under anomaly; prohibit autonomous escalation. Instrument operator interaction so leaders see dissenting sensor views rather than single "go/no-go" outputs (Stanley-Lockman, 2021; Federation of American Scientists, 2025; Yeung et al., 2021; Ppakter, 2024).
6. Build sovereign and allied options in compute and safety infrastructure. Invest in public compute (nationally and through alliances) and shared test ranges so that governments are not forced into single-vendor dependencies. Require portability, data residency, and open evaluation harnesses; fund safety professionals alongside sensors and satellites (Int'l Law Studies, 2021; Simmons-Edler et al., 2025; CSET, 2024; United Nations, 2024; Podar & Colijn, 2025; Toutoungi, 2025).
7. Prepare for deepfake-saturated crises. Adopt cryptographic content authentication for official communications; fund rapid-response attribution cells; and rehearse mixed-signal scenarios in exercises to avoid panic or paralysis when synthetic and kinetic events coincide (Twomey et al., 2023).
8. Bring nonaligned states into standard-setting by design. Offer adaptable artifacts—logging schemas, incident taxonomies, evaluation playbooks—to AU, ASEAN, and CELAC partners. This is a strategic move, not an act of charity. Shared artifacts create interop gravity and spread safer defaults beyond the transatlantic core (Podar & Colijn, 2025; Toutoungi, 2025; Devitt, 2021; Spektor, 2025).

Implementation caveats for major powers. Speed saves lives, but only when acceleration is paired with uncertainty-aware user interfaces and robust TEVV; otherwise, speed amplifies error and escalation risk (Stanley-Lockman, 2021; Department of Defense, 2022; Osoba & Welser, 2017). Some argue that adversaries will not self-bind, yet bounded autonomy can still reduce catastrophic risk—especially around NC3—and coalitions can establish de facto standards by making safety artifacts a condition of interoperability and access (Bruun & Bo, 2025; Ppakter, 2024; Int'l Law Studies, 2021; Simmons-Edler et al., 2025; United Nations, 2024). Others worry that detailed logs reveal sensitive methods, but tiered access and cryptographic controls can protect critical information while preserving accountability; aviation demonstrates that well-designed incident-sharing regimes can drive learning without exposing crown jewels (Greipl et al., 2024).

The framework maps directly onto the earlier diagnosis: compression (Sections 1-2) + unknowns (Section 3) + field practice (Section 4) + nuclear stakes (Section 5) + human judgment (Section 6) demand accountability (7.1), scalable standards (7.2), rigorous testing (7.4), NC3 safeguards (7.5), information-environment resilience (7.6), market-structure remedies (Section 8), and stack-aware geopolitics (Sections 9-10). Section 11 translates principles into specs; Section 12 provides the institutional machinery to test and enforce them. Policy is therefore not an appendix to technology; it is the blueprint that makes technology governable at speed.

CONCLUSION

Artificial intelligence is transforming national security less by raw capability than by compressing time, obscuring process, and reallocating agency. Across ISR, EW, targeting, and C2, learning systems move choices to the network edge, where milliseconds matter and small misclassifications can cascade at machine speed. In nuclear decision support, even advisory tools can bias human perception and reduce deliberation windows with existential consequences. Private ownership of critical stacks and multistack multipolarity further shift where power—and accountability—reside (Stanley-Lockman, 2021; Osoba & Welser, 2017; Federation of American Scientists, 2025; Yeung et al., 2021; Ppakter, 2024).

Three system-level findings follow. First, time has become a contested domain. Advantage accrues to actors who can operate inside adversary cycles without sacrificing reliability. Second, accountability is a design property: immutable logs, TEVV, and an incident regime are prerequisites for enforceable LOAC and credible deterrence at machine speed. Third, governance now travels through engineering: the actors who export evaluation harnesses, logging schemas, and interoperability defaults will write de facto norms long before diplomatic text is finalized (Bruun & Bo, 2025; Sati, 2023; Int'l Law Studies, 2021; Simmons-Edler et al., 2025; CSET, 2024; United Nations, 2024; Podar & Colijn, 2025; Toutoungi, 2025; Devitt, 2021; Spektor, 2025).

The policy path is therefore clear. Build an AI incident regime with safe-harbor reporting. Make TEVV contractual and pay-gated. Codify meaningful human control as a falsifiable specification (temporal/geographic bounds, target-profile limits, traceability, graceful degradation, uncertainty-aware UX). Harden NC3 with default-to-delay and dual-path adjudication. Pair compute export controls with calibrated training-run transparency. Invest in sovereign/allied computers and ranges to avoid sovereignty by subscription. Treat deepfake preparedness as command-integrity engineering. And bring nonaligned states into standard-setting by offering adaptable artifacts that scale safety beyond alliances.

The central choice is not automation versus restraint; it is structure versus drift. If we build the safety stack alongside the capability stack—logs, incidents, TEVV, ranges, human-factors design, NC3 safeguards, and inclusive standards—the machine-paced future can widen the space for judgment rather than collapse it. If we do not, an error at the edge can become a crisis at the core. Designing for time well used is the only sustainable strategy in a world where decisions increasingly happen at machine speed.

REFERENCES

Bajema, N., Gower, J., & Femia, F. (2022). *A handbook for nuclear decision-making and risk reduction in an era of technological complexity*. Council on Strategic Risks.

https://councilonstrategicrisks.org/wp-content/uploads/2022/12/NuclearTechnologicalComplexity-Dec22.pdf

Bo, M., Bruun, L., & Boulanin, V. (2022). *Retaining human responsibility in the development and use of autonomous weapon systems: On accountability for violations of international humanitarian law involving AWS*. https://doi.org/10.55163/ahbc1664

Bruun, L., & Bo, M. (2025). *Bias in military artificial intelligence and compliance with International Humanitarian Law*. https://doi.org/10.55163/NLWV5347

CSET. (2024). *AI safety and automation bias*. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/ai-safety-and-automation-bias/

Crootof, R. (2022). *AI and the actual IHL accountability gap*. Centre for International Governance Innovation. https://www.cigionline.org/articles/ai-and-the-actual-ihl-accountability-gap/

DARPA. (2020). *AlphaDogfight trials foreshadow future of human-machine symbiosis*. https://www.darpa.mil/news/2020/alphadogfight-trial

Department of Defense. (2022). *Summary of the Joint All-Domain Command & Control Strategy*. https://media.defense.gov/2022/Mar/17/2002958406/-1/-1/1/SUMMARY-OF-THE-JOIN-T-ALL-DOMAIN-COMMAND-AND-CONTROL-STRATEGY.PDF

Devitt, S. K. (2021). *Lethal autonomous weapons* [Preprint]. arXiv. https://arxiv.org/abs/2110.12935

Federation of American Scientists. (2025). *Artificial intelligence, and nuclear command, control, and communications: Current status and future risks*. https://fas.org/wp-content/uploads/2025/07/June2025_AKNC3_FAS.pdf

Greipl, A., Geneva Academy of International Humanitarian Law and Human Rights, & International Committee of the Red Cross. (2024). *Expert consultation report on AI and related technologies in military decision-making on the use of force in armed conflicts*. https://www.geneva-academy.ch/joomlatools-files/docman-files/Artificial%20Intelligence%20And%20Related%20Technologies%20In%20Military%20Decision-Making.pdf

Int'l Law Studies. (2021). Command accountability for AI weapon systems in the Law of Armed Conflict. *International Law Studies, 407*. https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=2958&context=ils

National Telecommunications and Information Administration. (2024). *Report on dual-use foundation models with widely available model weights*. U.S. Department of Commerce. https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

Nasu, H. (2021). *The Kargu-2 autonomous attack drone: Legal & ethical dimensions*. Lieber Institute West Point. https://lieber.westpoint.edu/kargu-2-autonomous-attack-drone-legal-ethical/

Newton, M. (2025). *How are drones changing war? The future of the battlefield*. Center for European Policy Analysis (CEPA). https://cepa.org/article/how-are-drones-changing-war-the-future-of-the-battlefield/

Osoba, O., & Welser, W., IV. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR1744.html

Podar, H., & Colijn, A. (2025). *Technical risks of (Lethal) autonomous weapons systems* [Preprint]. arXiv. http://arxiv.org/abs/2502.10174

Ppakter. (2024, September 4). *The risks and inefficiencies of AI systems in military targeting support*. Humanitarian Law & Policy Blog. https://blogs.icrc.org/law-and-policy/2024/09/04/the-risks-and-inefficacies-of-ai-systems-in-military-targeting-support/

Pusztaszeri, A., & Harding, E. (2025). *Technological evolution on the battlefield*. Center for Strategic and International Studies. https://www.csis.org/analysis/chapter-9-technological-evolution-battlefield

Risk Innovation Nexus. (2019). *Google Project Maven case study*. https://riskinnovation.org/wp-content/uploads/2019/10/Nexus_CaseStudy_Google_ProjectMaven_Final.pdf

Sati, M. C. (2023). The attributability of combatant status to military AI technologies under International Humanitarian Law. *Global Society, 38*(1), 122–138. https://doi.org/10.1080/13600826.2023.2251509

Simmons-Edler, R., Dong, J., Lushenko, P., Rajan, K., & Badman, R. P. (2025). *Military AI needs technically-informed regulation to safeguard AI research and its applications* [Preprint]. arXiv. https://arxiv.org/abs/2505.18371

Spektor, M. (2025, February 20). Rise of the nonaligned: Who wins in a multipolar world? *Foreign Affairs*. https://www.foreignaffairs.com/united-states/rise-nonaligned-multipolar-world-matias-spektor

Stanley-Lockman, Z. (2021). *Responsible and ethical military AI*. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/responsible-and-ethical-military-ai/

Toutoungi, A. (2025, July 31). *Ethics and regulation of AI in defence technology: Navigating the legal and moral landscape*. Taylor Wessing. https://www.taylorwessing.com/en/interface/2025/defence-tech/ethics-and-regulation-of-ai-in-defenc/e-technology

Trendsresearch.org. (2025). *Hallucinating machines: The trust issue in large language models*. https://trendsresearch.org/insight/hallucinating-machines-the-trust-issue-in-large-language-models/

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLOS ONE, 18*(10), e0291668. https://doi.org/10.1371/journal.pone.0291668

United Nations. (2024). *General and complete disarmament: Lethal autonomous weapons systems: Report of the Secretary General* (A/79/88). https://docs.un.org/en/A/79/88

U.S. Air Force. (2025). *Collaborative Combat Aircraft YFQ-42A takes to the air for flight testing*. https://www.af.mil/News/Article-Display/Article/4287627/

U.S. Government Accountability Office. (2022). *Extra Large Unmanned Undersea Vehicle (Orca) Program Assessment* (GAO-22-105974). https://www.gao.gov/assets/gao-22-105974.pdf

U.S. Navy. (2025). *Medium Unmanned Surface Vessel (MUSV)*. https://www.navy.mil/Resources/Fact-Files/Display/FactFiles/Article/4288073/

Wiese, L., & Langer, C. (2024). *Gaza, artificial intelligence, and kill lists*. Verfassungsblog. https://verfassungsblog.de/gaza-artificial-intelligence-and-kill-lists/

Yeung, D., Khan, I., Kalra, N., & Osoba, O. A. (2021). *Identifying systemic bias in the acquisition of machine learning decision aids for law enforcement applications*. RAND Corporation. https://www.rand.org/pubs/perspectives/PEA862-1.html