AUTONOMOUS WEAPONS AND MILITARY ARTIFICIAL INTELLIGENCE UNDER INTERNATIONAL LAW

The operational capability of an autonomous weapon system depends on tightly integrated hardware and software architecture designed to sense, navigate, and execute missions in dynamic physical environments.

Buğra Yalçın

Undergraduate Student, Political Science and International Relations, Abdullah Gül University / AVİM Trainee Program Participant

INTRODUCTION

The rapid integration of Artificial Intelligence (AI) and autonomous technologies into military systems marks one of the most transformative shifts in the history of warfare. Modern armed forces are shifting from remotely piloted platforms to autonomous systems with developed technologies in data analysis, identifying target with minimal or no human intervention. While these technologies offer significant tactical advantages, they also present profound challenges to established international legal and moral framework.

At the center of this debate is the tension between autonomous decision-making and the requirements of International Humanitarian Law. Core principles such as distinction, proportionality, and precautions were developed around the cogni-



tive capacities, moral choices, and subjective judgment of human commanders. Delegating critical functions of targeting to opaque algorithms raises urgent questions: Can machine learning models reliably distinguish between active combatants, civilians, and surrendering soldiers? Who bears legal responsibility when an autonomous weapon causes unintended civilian casualties? How can the international community establish effective oversight before geopolitical rivalries trigger an unconstrained algorithmic arms race?

This report provides a comprehensive analysis of military AI and autonomous weapon systems across three interrelated dimensions:

- **Chapter 1:** Technical Architecture, Operational Dynamics, and System Risks
- **Chapter 2:** Autonomous Weapons under International Law and Criminal Responsibility
- **Chapter 3:** Future Perspectives, Oversight, and Global Regulatory Efforts and Practical Applications

CHAPTER 1:

TECHNICAL ARCHITECTURE, OPERATIONAL DYNAMICS, AND SYSTEM RISKS

1.1. Historical Evolution and Technological Transformation

The modernization of military capabilities has been fundamentally driven by the dual objectives of removing human soldiers from immediate physical danger and accelerating the tempo of battlefield decision-making. Over recent decades, military systems have evolved from manually operated platforms and remotely piloted drones toward self-contained architectures powered by machine learning and autonomous software. This transformation spans both munitions technology -seen in the rise of loitering munitions capable of autonomous loitering and terminal homing- and the broader evolution of defense artificial intelligence. By integrating advanced data analytics into Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR) networks, armed forces have restructured targeting pipelines, introducing high-speed predictive models that fundamentally reshape the tactical and doctrinal character of contemporary armed conflict.

1.1.1. Transition from Remotely Operated Systems to Full Autonomy

The evolution of military technologies has shifted in light of the goal of reducing human physical presence on the battlefield and achieving faster decision-making processes. Modern operational challenges have pushed military systems to develop toward autonomous technologies to create an independent and rapid decision-making mechanism. Over time, human local planners have increasingly been replaced or enhanced by deep learning and reinforcement learning frameworks.

This paradigm shift is visible in munitions technology, specifically in the transition from traditional guided cruise missiles to “loitering munitions” (LM). This shift completely changed the understanding of war, as LMs can search for targets, identify them, and execute attacks autonomously. The transition of these technologies and the understanding of war into lethal autonomous systems have been made possible by integrated decision pipelines that render real-time human approval seemingly redundant before an attack.

1.1.2. The Evolution of Artificial Intelligence in Military Technologies

The development of artificial intelligence in the defense sector has followed three distinct historical waves:

- **First Wave (Rule Based Expert Systems):** AI systems operated strictly on predefined logical rules and «if-then» structures, but could not adapt to the chaotic conditions of real battlefields.
- **Second Wave (Statistical Machine Learning):** Rather than following rigid rules, al-

gorithms learned patterns directly from data, achieving practical success.

- **Third Wave (Deep Learning and Neuromorphic Systems):** Driven by deep neural networks and massive data availability, such models achieve near-human accuracy in target detection and battlespace perception.

Integrating AI into military command structures has fundamentally reshaped the C4ISR framework. By using Big Data based advanced analytics, AI-driven systems automate the entire intelligence cycle, offering fast tactical assessments and recommended action plans to commanders.

The drive for algorithmic speed has also introduced controversial concepts into modern operational doctrines, specifically "predictive violence" and "algorithmic pre-emptive strikes". In contrast to traditional laws of armed conflict, predictive machine learning models analyze massive historical datasets to estimate the likelihood that an individual will pose a future threat. This practice has led to what critics term "death by metadata", where targeting decisions depend on calculated risk scores rather than clear verification of a hostile act.

1.2. Technical Architecture and Core Technologies

The operational capability of an autonomous weapon system depends on tightly integrated hardware and software architecture designed to sense, navigate, and execute missions in dynamic physical environments. At the foundational level, platforms utilize a multi-layer robotic navigation stack -combining perception, mapping, localization, and hierarchical path planning- supported by multi-sensor data fusion to maintain situational awareness. At the tactical execution level, these physical platforms are paired with advanced decision-making and targeting algorithms, including Hierarchical Reinforcement Learning (HRL) for close-in weapon defense and AI Decision Support Systems (AI DSS) for automated target generation. However, while these computational pipelines provide rapid processing capabilities, they simultaneously introduce severe cognitive vulnerabilities, such as automation bias and algorithm aversion, which impact human oversight during high-stress combat operations.

1.2.1. Fundamental Components of Autonomous Weapon Systems

To operate independently in dynamic physical battlefields, autonomous weapon systems require an integrated navigation pack which includes five core components:

- **Perception Layer:** Gathers raw environmental data using physical sensors integrated on the platform.
- **Mapping Layer:** Constructs spatial representations of the area using Simultaneous Localization and Mapping (SLAM) algorithms.
- **Localization Layer:** Continuously calculates the platform's exact position and orientation relative to the map.
- **Global Planner:** Uses graph search algorithms to compute the shortest, collision-free global path from the current position to the mission target.
- **Local Planner:** Generates high-frequency motor commands to avoid unmapped obstacles while keeping the platform aligned with the global path.

A system's ability to interpret its surroundings accurately depends heavily on multi-sensor fusion technology across architectural fusion levels that culminate in all sensor information being combined using statistical voting.

1.2.2. Sensing, Decision-Making, and Targeting Algorithms

In modern warfare, engagement speeds often exceed human physiological capabilities. This is particularly evident in Close-In Weapon Systems (CIWS) deployed to protect naval vessels and ground bases from saturation attacks by drone



swarms. With that protection circle, military bases have attained a new level of security.

Because conventional control systems struggle to track multiple fast-moving targets executing erratic maneuvers, advanced CIWS platforms increasingly employ Hierarchical Reinforcement Learning (HRL) architectures:

- **High-Level Policy (Meta-Policy):** Evaluates incoming threats based on range, velocity vectors, and approach angles to decide which target must be engaged first.
- **Low-Level Policy (Sub-Policy):** Calculates the lead-angle targeting solutions, drives the turret motors, and applies residual control adjustments to correct for ballistic projectile dispersion.

This distributed coordination is further modeled using Multi-Living Agent (MLA) theory and two-stage queuing networks. In integrated missile-gun defense installations, individual search radars, tracking sensors, and firing units are treated as independent collaborative agents. If, for example, a specific launcher experiences a mechanical failure, runs out of ammunition, or becomes saturated, adjacent units dynamically reallocate targets across the defense network to prevent defensive gaps.

Beyond autonomous air defense, algorithmic targeting is expanding through AI Decision Support Systems (AI DSS) designed for ground and air operations. Systems such as the United States (US) military's Maven Smart System (MSS) and the Israeli military's Habsora (The Gospel) and Lavender systems process high-resolution video and intelligence databases to generate automated target lists. While these capabilities grant states immense tactical superiority and rapid target generation, they simultaneously create a fundamental conflict of interest by reducing human lives to probabilistic statistical risk scores.

Furthermore, integrating AI decision support introduces two significant cognitive risks for human operators:

- **Automation Bias:** Occurs when human operators place unjustified trust in algorithmic recommendations, reducing their own



critical vigilance and failing to spot errors produced by the software. Field reports regarding the Lavender system indicate that human reviewers spent as little as 20 seconds verifying each algorithmic target proposal, functioning largely as rubber stamps for automated kill lists. In practice, this turns the theoretical requirement of human oversight into a mere procedural illusion, allowing states to maximize operational tempo while shifting the catastrophic risk of statistical false positives directly onto innocent civilians.

- **Algorithm Aversion:** The opposite psychological tendency, where operators completely reject algorithmic recommendations after observing a single software error, even if the algorithm statistically outperforms human analysts.

Empirical studies show that specialized military training mitigates automation bias under baseline conditions; however, when subjected to intense time pressure and high cognitive workloads, even experienced operators tend to uncritically accept algorithmic suggestions. This structural vulnerability demonstrates that as states delegate critical identification tasks to algorithms, the absence of absolute reliability inevitably exposes civilian populations to lethal misjudgments under high-stress combat environments.

1.3. Levels of Autonomy and Classification

To evaluate the operational, ethical, and legal boundaries of emerging military technologies, international legal and defense frameworks categorize systems according to the degree of human involvement in the decision-making cycle. Central to this analysis is the concept of the Observe, Orient, Decide, Act (OODA) loop, which distinguishes between systems that keep humans “in the loop” (HITL), “on the loop” (HOTL), or entirely “out of the loop” (HOOTL). In multilateral disarmament and legal forums, establishing clear taxonomies is essential to differentiate between ordinary automated navigation aids and Lethal Autonomous

Weapon Systems (LAWS) that possess autonomy in their critical functions of target selection and engagement. As modern combat increasingly experiences "decision-speed asymmetry", defining these operational tiers becomes vital for analyzing whether human control can be meaningfully maintained or if military necessity will drive warfare toward unconstrained algorithmic conflict.

Empirical studies show that specialized military training mitigates automation bias under baseline conditions; however, when subjected to intense time pressure and high cognitive workloads, even experienced operators tend to uncritically accept algorithmic suggestions.

1.3.1. Loop Concepts: HITL, HOTL, and HOOTL

The classification of autonomous weapon systems depends on the location and role of human operators within the decision (OODA) loop. International legal and military frameworks distinguish three main configurations:

- **Human-in-the-Loop (HITL):** The weapon system can search for, detect, and track potential targets automatically, but the decision to apply lethal force requires explicit, active human command. Standard remotely piloted drones such as the MQ-9 Reaper belong to this category.
- **Human-on-the-Loop (HOTL):** The system has the software authorization to search, select, and fire upon targets autonomously, while a human supervisor monitors operations with the technical ability to interrupt, veto, or abort the engagement. Automated point-defense systems like Phalanx CIWS or Patriot batteries often operate in this supervisory mode due to tight engagement time constraints.
- **Human-out-of-the-Loop (HOOTL):** The system executes the entire cycle of target search, classification, engagement, and post-strike assessment without any human intervention or real-time oversight. This configuration represents a fully autonomous weapon system (LAWS).

Although these definitions appear straightforward, they reveal what philosophers term the "autonomous machine fallacy". Equipping a computer with self-learning software does not grant it genuine moral agency. Human beings operate as "moral wrestlers", weighing ethical duties, compassion, and legal norms during combat; machines, by contrast, merely optimize mathematical objective functions.

Consequently, deploying fire-and-forget, non-recallable autonomous systems that cannot be reassessed or safely recalled after launch creates a dangerous moral and legal vacuum on the battlefield. This vacuum directly threatens civilian lives, as unrecallable algorithms operating in dynamic environments cannot exercise human restraint or reassess changing humanitarian conditions on the ground.

1.3.2. Distinction Between Lethal Autonomous Weapon Systems (LAWS) and Semi-Autonomous Systems

In international humanitarian law and diplomatic disarmament forums, the core debate centers on the difference between semi-autonomous systems and fully autonomous LAWS. A platform that flies, navigates, or lands autonomously is not necessarily an autonomous weapon; the defining legal factor is autonomy in its critical functions, as in, selecting and applying force to human targets without human intervention.

To establish regulatory clarity, Laiq (2026) proposes a five-tiered taxonomy of military autonomy:

- **Category 1 (Zero Autonomy):** Human operator only (manual artillery, unguided rockets) – Fully conventional; strictly compliant with international humanitarian law (IHL).

Deploying fire-and-forget, non-recallable autonomous systems that cannot be reassessed or safely recalled after launch creates a dangerous moral and legal vacuum on the battlefield.

- **Category 2 (Assisted Autonomy):** Human with sensor aids (laser-guided munitions, thermal scopes) – Semi-autonomous targeting; human aims and fires.
- **Category 3 (Semi-Autonomous / HITL):** Algorithm with human command (Bayraktar TB2, MQ-9 Reaper) – Algorithm detects; human gives final strike command.
- **Category 4 (Supervised Autonomy / HOTL):** Algorithm with human veto (Phalanx CIWS, Iron Dome, Harop) – Machine fires automatically; human can abort.
- **Category 5 (Full Autonomy / HOOTL):** Autonomous algorithm (autonomous Kargu-2, autonomous swarms) – LAWS status; primary focus of international ban proposals.

Recent conflicts highlight the operational deployment of higher-tier systems. In the 2021 United Nations (UN) Panel of Experts report on Libya, an autonomous Kargu-2 loitering drone was documented as having hunted and attacked retreating forces without active communication with an operator. Similarly, during the 2020 Karabakh War, Harop loitering munitions operated in autonomous homing modes against active radar emissions.

When algorithms make targeting choices in microseconds, opposing forces are pressured to deploy automated counter-systems with equal speed. This dynamic leads to "decision-speed asymmetry", threatening to pull human control entirely out of the loop and shifting warfare toward high-speed, unpredictable algorithmic conflict, alternatively known as "hyperwar". This systemic pressure creates a dangerous geopolitical dilemma; in their quest to maintain strategic parity, states are incentivized to bypass human deliberative safeguards, accepting the inherent unreliability of autonomous platforms and heightening the danger of widespread civilian harm in contested zones.

1.4. System Reliability, Limitations, and Risks

Despite the tactical advantages offered by AI, the deployment of autonomous systems in real-world combat operations is constrained by critical technical limitations and systemic failure modes. Deep learning architectures are susceptible to algorithmic bias across their data collection and development lifecycles, often suffering from severe simulation-to-reality transfer gaps ("Sim-to-Real gap") and mathematical opacity known as the "Black Box" problem. This inherent opacity generates instances of "artificial stupidity", where machine learning models produce unpredictable and catastrophic misclassifications that human operators cannot readily explain or foresee. Furthermore, autonomous platforms operate in contested electronic environments where they are vulnerable to Adversarial Machine Learning (AML) techniques, including evasion perturbations and training data poisoning attacks. Together, these intrinsic model errors and external cyber vulnerabilities make autonomous systems highly unpredictable on the battlefield, creating significant risks for civilian protection and mission safety.

1.4.1. Algorithmic Error, Black Box Problem, and Bias

The deployment of machine learning in combat systems faces fundamental technical obstacles, primarily algorithmic bias and model opacity. Rather than being a simple coding error, algorithmic bias in military AI is a socio-technical issue that spans three distinct lifecycle stages:

1. **Data Collection Stage:** Deep learning models reflect the limitations of their training data. Because operational military data is highly classified and scarce, models are often trained on narrow geographic datasets (such as arid, open desert environments) that fail

to represent dense urban areas or diverse global terrains. When exposed to unfamiliar background environments, object detectors can make severe classification errors. Furthermore, training systems on demographic proxies (such as gender, age, or clothing styles) risks systematically misclassifying all military-aged civilian males in a conflict zone as enemy combatants.

2. **Design and Development Stage:** Mathematical loss functions inevitably encode engineering assumptions. For example, in robotics, training policies in idealized, symmetric simulated corridors causes severe transfer problems (Sim-to-Real gap). A policy trained only on balanced corridors experiences observation aliasing and overfitting; once deployed in the physical world, slight sensor noise causes the robot to drift, and lacking asymmetric corrective training, it cannot recover.

3. **Operational Deployment Stage:** Under extreme battlefield stress, human personnel frequently accept flawed AI recommendations without verification, turning latent software bias into civilian casualties.

These failure modes are exacerbated by the "Black Box" problem inherent to deep neural networks. Because a deep network comprises millions of interconnected data across hidden layers, developers cannot trace the exact mathematical chain of reasoning that led the system to classify an object as a military target.

This opacity creates "artificial stupidity", where algorithms make catastrophic errors that no human would commit. Computer vision researchers have demonstrated that slight surface texture changes can cause an image classifier to identify an animal toy as a rifle with 99% statistical confidence. Similarly, placing small strips of reflective tape on a standard STOP sign causes autonomous driving vision models to misclassify it as a high-speed limit sign. Without robust Explainable AI (XAI) verification methods, deploying opaque neural networks in lethal platforms creates severe operational and humanitarian risks. This structural opacity proves that states, in their pursuit of automated tactical superiority, deploy systems whose zero-error reliability can never be guaranteed, thereby shifting the fatal consequences of algorithmic unreliability directly onto non-combatants.

1.4.2. Cyber Vulnerabilities and Unpredictability in the Battlespace

Autonomous weapon systems are vulnerable to targeted cyber and electronic manipulation. The study of these vulnerabilities is known as Adversarial Machine Learning (AML), which investigates methods to deceive, manipulate, or degrade machine learning models.

In cyber warfare, attacks against machine learning architectures fall into three main categories:

- **Evasion Attacks:** Executed during the inference phase while the system is actively deployed. An adversary adds calculated mathematical perturbations to an object that are invisible to human observers but cause the neural network to misclassify the input with high confidence. Algorithms such as the Fast Gradient Sign Method (FGSM), DeepFool, and Carlini-Wagner can alter an armored fighting vehicle's visual signature so that targeting software classifies it as a civilian school bus -or conversely, manipulate civilian vehicles to appear as high-priority military targets.

- **Poisoning Attacks:** Executed during the training phase. If an autonomous system trains on open-source, un-curated, or intercepted field data, an attacker can inject mali-

Despite the tactical advantages offered by AI, the deployment of autonomous systems in real-world combat operations is constrained by critical technical limitations and systemic failure modes.



cious training samples to distort the decision boundary. In a "boiling frog" attack, the adversary introduces small amounts of misleading data over months; leading the system to gradually accommodate this noise, eventually blinding itself to specific enemy vehicle signatures.

Existing IHL norms were designed around human judgment, moral choices, and human conscience. For this reason, evaluating how autonomous systems can comply with these core principles requires detailed legal and technical analysis, highlighting the acute conflict of interest between a state's drive for autonomous capability and its international humanitarian obligations.

- **Model Inversion and Extraction Attacks:** An adversary repeatedly queries an AI system's input-output interface to reverse-engineer its internal parameters and reconstruct sensitive training data.

A major operational threat is the "transferability property" of adversarial perturbations. An adversarial pattern generated to fool one specific neural network model will frequently deceive completely different model architectures trained on separate datasets. Consequently, an adversary who develops an adversarial camouflage pattern in a private lab can apply it to ground assets, successfully blinding a wide range of unknown autonomous sensor platforms on the battlefield.

Standard software defenses such as gradient masking or defensive distillation do not provide reliable security against black-box transfer attacks. Susceptible to both intrinsic classification failures and deliberate cyber manipulation, fully autonomous weapons remain highly unpredictable systems

when deployed in contested combat environments. Because states are driven by geopolitical rivalry to deploy these systems despite known vulnerabilities, the battlefield unpredictability of manipulated models inevitably endangers civilian structures and non-combatants situated in operational theaters.

CHAPTER 2:

AUTONOMOUS WEAPONS UNDER INTERNATIONAL LAW AND CRIMINAL RESPONSIBILITY

2.1. Assessment under International Humanitarian Law (IHL)

International Humanitarian Law is a body of rules that establishes that the means and methods of warfare are not unlimited and that persons who do not take part in hostilities must be protected. Because there is currently no specific and legally binding international treaty that specifically governs AI-supported autonomous weapon systems, existing IHL rules must be applied directly to these emerging technologies. Existing IHL norms were designed around human judgment, moral choices, and human conscience. For this reason, evaluating how autonomous systems can comply with these core principles requires detailed legal and technical analysis, highlighting the acute conflict of interest between a state's drive for autonomous capability and its international humanitarian obligations.

2.1.1. The Principle of Distinction

The principle of distinction is a cornerstone of IHL and is codified in Articles 48, 51(2), and 52(2) of the 1977 Additional Protocol I (AP I) to the Geneva Conventions. Under this rule, parties to a conflict must always distinguish between the civilian population and combatants, as well as between civilian objects and military objectives, directing their operations solely against military objectives.

Compliance of autonomous weapons with the principle of distinction requires a three-stage technical and cognitive process:

1. **Observation Capability:** Gathering environmental information using advanced sensors such as LIDAR, radar, thermal imaging, and optical cameras. For instance, mapping drones developed by Raytheon and Exyn Technologies collect up to 300,000 data points per second to create real-time 3D models of underground tunnels and complex indoor areas, observing physical spaces faster than human soldiers.

2. **Recognition Capability:** Matching collected sensor data against stored target databases. Systems such as Patriot One's «PatScan" detect and classify specific weapon types with a 94.7% accuracy rate, identifying physical threats more rapidly and consistently than human security guards.

3. **Contextual Judgment Capability:** Determining whether a person is legally a legitimate target based on the surrounding context rather than purely on visual features.

Although AI systems excel at observation and pattern recognition, they face significant limitations in contextual judgment. Under IHL, a uniformed soldier who is wounded, unconscious, or clearly expressing an intention to surrender is considered *hors de combat* (out of action) and cannot be targeted (AP I, Article 41). Recognizing gestures of surrender -such as raising hands, dropping weapons, or waving a white flag- requires evaluating subjective human intent.

Just as agricultural robots struggle when insect bites or weather changes alter leaf colors and shapes, military computer vision algorithms can misclassify targets when unexpected changes occur in chaotic combat environments. Furthermore, under Article 50(1) of AP I, if there is doubt regarding the civilian status of a person, that person must be presumed to be a civilian. Programming an algorithm to evaluate the legal "threshold of doubt" and hold fire in dynamic combat situations remains an unresolved technical and legal challenge. Because probabilistic classifiers operate on statistical likelihood rather than legal doubt, deploying autonomous weapons in contested civilian areas risks systematically erod-

Although AI systems excel at observation and pattern recognition, they face significant limitations in contextual judgment.

ing the fundamental presumption of civilian protection.

2.1.2. The Principle of Proportionality

The principle of proportionality is codified in Article 51(5)(b) of AP I and in Article 8(2)(b)(iv) of the Rome Statute of the International Criminal Court (ICC). It prohibits attacks against legitimate military objectives if the attack is expected to cause incidental loss of civilian life, injury to civilians, or damage to civilian objects that would be excessive in relation to the concrete and direct military advantage anticipated.

Evaluating proportionality involves a three-step test:

- **Collateral Damage Estimation (CDE):** Calculating expected incidental harm based on weapon type, blast radius, building materials, and civilian population density. Because CDE relies heavily on physics and blast models, software can perform these calculations with high precision.
- **Anticipated Military Advantage (AMA):** Assessing the direct tactical or operational gain from an attack, such as capturing ground or destroying an adversary's combat capability.
- **Determination of Excessiveness:** Comparing the estimated collateral damage with the expected military advantage to decide if the expected harm is excessive.

The core issue in proportionality is the problem of "incommensurability". Civilian lives and military advantages represent fundamentally different values that cannot be converted into a single mathematical metric. International tribunals assess this balance through the subjective judgment and good faith of a "reasonable military commander".

To address this challenge in autonomous systems, Zając (2023) outlines three operational solutions:

1. **Operations in Civilian-Free Environments:** Restricting autonomous weapons to areas where civilians are absent (such as open waters, subsea domains, deserts, or high altitudes), removing the need for real-time proportionality calculations.
2. **Pre-Authorized Targeted Strikes:** Requiring human commanders to evaluate the target's value and set a strict ceiling for acceptable collateral damage in advance, allowing the weapon to strike only if its sensor-based CDE remains below that pre-approved threshold.
3. **The "Price Tag" Approach in Attritional Warfare:** Allowing commanders to assign pre-calculated allowable collateral damage values to specific classes of enemy equipment (such as tanks or radar installations) within defined geographic and temporal boundaries.

However, substituting the contextual judgment of a reasonable commander with pre-set algorithmic "price tags" risks standardizing civilian casualties into an acceptable mathematical margin of error, fundamentally undermining the qualitative safeguards that IHL affords to civilian populations.

2.1.3. Military Necessity, Unnecessary Suffering, and the Pillars Theory

The legal limits of warfare trace back to the 1868 St. Petersburg Declaration and the 1907 Hague Regulations. These instruments establish that the only legitimate objective of war is to weaken the military forces of the enemy, and that using weapons of a nature to cause superfluous injury or unnecessary suffering is strictly prohibited (codified in AP I, Article 35(2)).

Winter (2020, 2022) argues that in IHL theory, humaneness and military necessity are not separate "legal principles," but rather the foundational "pillars" of the entire legal framework:

- **Application of Legal Principles (Ronald Dworkin's Model):** Genuine legal principles carry varying weights, compete against each other, and can override specific positive rules when they conflict.
- **The Balancing Nature of IHL:** Military necessity and humaneness operate as counterbalancing forces. Military necessity cannot override treaty prohibitions (such as the absolute ban on harvesting organs from prisoners of war).

These two pillars are embedded directly within positive rules such as distinction, proportionality, and precautions. If an autonomous weapon is engineered to comply with these positive rules, the broader demands of military necessity and humaneness are satisfied simultaneously. Yet, when states prioritize military necessity and algorithmic speed over humaneness without verifiable technical safeguards, this delicate normative equilibrium is broken to the detriment of non-combatants.

2.1.4. The Martens Clause and Human Dignity

Codified in the 1899 and 1907 Hague Conventions and reaffirmed in Article 1(2) of AP I, the Martens Clause provides that in cases not covered by specific treaty provisions, civilians and combatants remain under the protection of the "principles of humanity and the dictates of public conscience".

Ethical and legal debates on autonomous weapons and human dignity generally split into two main perspectives:

- **Prohibitionism:** Scholars such as Peter Asaro and Robert Sparrow argue that allowing



Civilian lives and military advantages represent fundamentally different values that cannot be converted into a single mathematical metric.

an algorithm to make lethal decisions treats human beings like inanimate targets or vermins, violating human dignity and requiring an outright ban on LAWS.

- **Weak Restrictionism:** Leonard Kahn distinguishes between «basic human dignity» (the right of non-combatants to live and not be targeted) and «non-basic human dignity» (concerns regarding combatants being targeted by machines). Under Kahn's model, if an advanced autonomous system can distinguish between combatants and non-combatants more reliably than human soldiers, using that system protects basic human dignity and becomes morally permissible or even obligatory.

However, because algorithmic certainty remains technically unattainable, the premise of Weak Restrictionism collapses in practice, leaving civilian dignity and basic life subject to the dictate of opaque and fallible automated code.

2.1.5. Article 36 of Additional Protocol I (New Weapon Reviews)

Article 36 of AP I of the 1949 Geneva Conventions requires states to determine, during the study, development, acquisition, or adoption of any new weapon, means, or method of warfare, whether its employment would be prohibited by international law.

- **Historical Roots:** Natalia Jevglevskaja (2014–2015) notes that Article 36 was introduced during the 1974–1977 Diplomatic Conference primarily by Germany and the United Kingdom, building upon Germany's 1961 military manual (ZDv 15/10) which required new weapons to comply with international regulations.
- **Compliance-by-Design:** Bérénice Boutin (2023) points out that because deep learning algorithms are opaque, post-deployment reviews are insufficient; under Common Article 1 of the Geneva Conventions (the duty to respect and ensure respect for IHL), legal constraints must be integrated directly into the software development, training datasets, and testing stages.
- **Limitations as a Regulatory Stopgap:** Tsybulenko and Kajander (2022) argue that because states often keep Article 36 review results classified, this mechanism lacks international transparency and cannot fully substitute for a dedicated, binding treaty on autonomous weapons. This structural secrecy allows states to field powerful AI weaponry under the guise of classified compliance, side-stepping public scrutiny and international accountability while civilian safety remains unverified.

2.1.6. The Principle of Precautions in Attack

Codified in Article 57 of AP I, the principle of precautions in attack requires planners and operators to take all possible measures to avoid or minimize civilian harm. Precautionary obligations include four main components:

1. **Target Verification:** Taking all feasible steps to verify that objectives are legitimate military targets.
2. **Mitigation of Means and Methods:** Selecting weapon types, attack angles, timing, and fuze delays to minimize collateral damage. Standardized frameworks like the US Collateral Damage Estimation Methodology (CDEM) show that these evaluations can be calculated systematically.
3. **Suspension or Cancellation:** Halting an attack if it becomes apparent that the target is non-military or that the strike will violate proportionality.
4. **Advance Warnings:** Providing effective

warnings to civilian populations when circumstances permit. Because autonomous platforms carry no risk to human pilots, they can operate closer to hostile areas to deliver warnings (via broadcasts or sirens) without placing military personnel in danger, raising the standard of what is practically feasible in precautions.

Nevertheless, if an autonomous system lacks continuous human verification, the legal obligation to abort an attack in real-time upon unexpected civilian movement becomes compromised, rendering precautionary obligations secondary to automated mission completion.

2.2. Individual Criminal Responsibility and Command Responsibility in International Criminal Law

Because AI lacks legal personhood and cannot form criminal intent (*mens rea*), criminal liability cannot be attached to algorithms. Instead, accountability rests exclusively with human actors. This section examines how *actus reus*, *mens rea*, command responsibility under Article 28 of the Rome Statute, and manufacturer complicity apply to military AI to prevent a culture of impunity.

2.2.1. Material and Mental Elements (*Actus Reus* and *Mens Rea*)

Under Article 30 of the ICC Rome Statute, individual criminal responsibility requires proving both the physical conduct (*actus reus*) and the mental state (*mens rea* - intent and knowledge). In the context of military AI, the core challenge in assessing criminal liability stems from the fact that an AI system is legally a weapon or a tool rather than a natural or legal person in the sense that it completely lacks moral agency, intentionality, and the capacity to form *mens rea*. Consequently, an unlawful attack resulting in civilian harm (*actus reus*) cannot generate criminal liability for the algorithm itself. Instead, criminal responsibility rests exclusively on the human developers, commanders, and operators who design, authorize, and deploy these systems.

- **Legal Status of AI:** Autonomous systems lack moral agency and legal personhood; they cannot form intent and cannot be punished under criminal law.
- **Mistake of Fact:** Under Article 32 of the Rome Statute, if a human commander honestly and reasonably believed based on an algorithm's output that an object was a legitimate military target, the lack of *mens rea* may exclude criminal liability for an unintentional attack on civilians.
- **Pattern Evidence and Oblique Intent:** Milanovic (2025/2026) explains that if an AI system repeatedly misidentifies civilian infrastructure and a commander continues to deploy it without correction, intent can be inferred from patterns of repeated failure and reckless disregard.
- **Innocent Agency:** Akkuş (2023) highlights that if an operator intentionally uses an autonomous weapon to execute an unlawful strike, the human remains liable as an indirect perpetrator using the machine as an instrument under Article 25(3)(a) of the Rome Statute.

This doctrinal framework exposes a severe gap; when civilian deaths result from unexpected algorithmic misclassifications, the invocation of "mistake of fact" based on blind trust in algorithmic output risks creating a culture of impunity where no individual is held criminally accountable.

2.2.2. Command Responsibility under Article 28 of the Rome Statute

Command responsibility holds military commanders and civilian superiors criminally liable for

Because AI lacks legal personhood and cannot form criminal intent *(mens rea)*, criminal liability cannot be attached to algorithms. Instead, accountability rests exclusively with human actors.

Command responsibility holds military commanders and civilian superiors criminally liable for crimes committed by subordinates under their effective command and control, when the superior knew or should have known about the crimes and failed to take necessary and reasonable measures to prevent or punish them.

crimes committed by subordinates under their effective command and control, when the superior knew or should have known about the crimes and failed to take necessary and reasonable measures to prevent or punish them.

Spadaro (2023), in "A Weapon is No Subordinate," demonstrates why command responsibility cannot be applied directly to autonomous machines:

- **Nature of Effective Control:** Command responsibility applies to hierarchical relationships between human superiors and human subordinates, not to control over technical equipment.
- **Inability to Punish or Discipline:** A commander cannot discipline, court-martial, or punish a software algorithm.
- **Derivative Nature of Liability:** Command responsibility requires an underlying criminal offense committed by a subordinate because an algorithm cannot commit a crime, and derivative liability cannot be established solely through the machine's actions.

Command responsibility applies to the relationship between the human commander and the human subordinates who program, authorize, or deploy autonomous weapons. This structural limitation ensures that if an autonomous system behaves unpredictably and harms civilians, traditional



command responsibility doctrine struggles to attach liability to the commander for the machine's autonomous acts.

2.2.3. Responsibility of Civilian Superiors, Programmers, and Manufacturers

The involvement of multiple private and technical actors in developing autonomous systems extends potential liability to software engineers and corporate executives.

- **Aiding and Abetting:** Under Article 25(3)(c) and (d) of the Rome Statute, individuals who provide the means to commit a crime with knowledge of the group's intent are criminally liable. If software engineers or defense contractors knowingly provide targeting software programmed to carry out indiscriminate attacks, they may be prosecuted for complicity.



- **Historical and Corporate Precedents:** Cases ranging from the post-Second Word War Zyklon B trials to the recent the Lafarge company prosecution in France confirm that corporate actors who provide technical means for international crimes with knowledge of their use face individual criminal responsibility.
- **2014 UN Arms Trade Treaty (ATT):** Articles 6 and 7 of the ATT prohibit states from authorizing arms transfers (including AI software components) if there is knowledge or an overriding risk that they will be used to commit war crimes.

Despite these provisions, diffusing responsibility across complex supply chains and commercial software developers makes proving individual *mens rea* exceptionally difficult, leaving civilian victims without direct avenues for justice.

2.3. The Responsibility Gap and State Responsibility

The unpredictability of machine learning creates a "responsibility gap," where human decision-makers may use algorithmic opacity to evade personal blame. To prevent impunity when individual criminal intent cannot be proven, this section analyzes objective state responsibility under the UN International Law Commission (ILC) Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA), ensuring that states remain directly accountable for unlawful civilian harm caused by autonomous weapons.

2.3.1. Debates on Agency and Cognitive Factors

The self-learning and unpredictable behaviors of machine learning models have led some theorists to argue that an accountability gap arises when neither the developer nor the user can foresee specific system outputs.

Oimann and Salatino (2025) analyze the psychological and empirical dimensions of this issue:

- **Reduced Sense of Agency (SoA):** Experimental fMRI and EEG studies show that placing intermediaries (human or automated agents) between commanders and actions reduces the commander's implicit sense of agency and empathetic responses.
- **Strategic Moral Scapegoating:** Commanders might use the technical opacity of AI systems to avoid personal blame, pointing to algorithmic unpredictability to evade accountability. By treating the software as an autonomous scapegoat, states and military hierarchies can utilize the destructive power of autonomous systems while shielding decision-makers from the legal repercussions of civilian casualties.
- **Limits of the Yamashita Precedent:** Relying on strict command liability models (such as the Yamashita ruling after the Second Word War) in opaque autonomous systems risks imposing unjust liability on commanders for software behaviors they could neither foresee nor control.

2.3.2. State Responsibility (ARSIWA), Mistake of Fact, and Object Agency

Unlike international criminal law, which requires individual *mens rea*, the law of State Responsibility (codified in ARSIWA) is an objective liability regime. State responsibility requires two core elements (ARSIWA, Article 2): 1) Attribution of conduct to the state under international law, 2) A breach of an international obligation of that state.

- **Attribution via State Organs:** When a commander deploys an autonomous weapon that commits an unlawful attack, that conduct is attributable to the state under Article 4 of ARSIWA (even if the commander exceeded authority or disobeyed orders, pursuant to Article 7).
- **Objects as State Agents (De Lege Ferenda):** Magdalena Pacholska (2023) proposes that if fully autonomous systems (FAWS) operating without human intervention are ever fielded, international law should draw on the broad definition of «agent» from the ICJ's *Reparation for Injuries* advisory opinion. Under this approach, the weapon itself could be treated as an agent of the state, attributing its unlawful acts directly to the fielding state without requiring fault from individual operators.
- **Phased Regulation Model:** Nicholas Cromie (2026) suggests a middle path between blanket bans and unregulated use: testing fixed, defensive autonomous systems in low-risk environments (such as static counter-drone systems) to empirically prove IHL compliance before considering more complex mobile offensive applications.

The unpredictability of machine learning creates a "responsibility gap," where human decision-makers may use algorithmic opacity to evade personal blame.

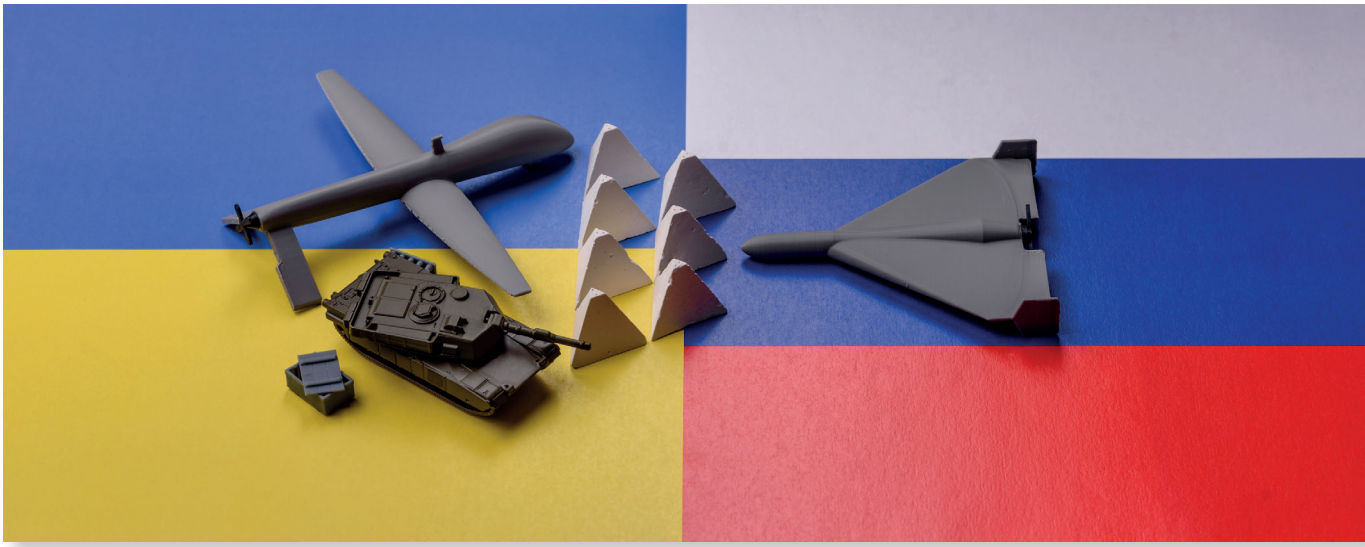
Given the fundamental uncertainty of autonomous software and the difficulty of establishing individual criminal guilt, robust and objective state responsibility under ARSIWA serves as the indispensable legal mechanism to prevent states from externalizing the catastrophic risks of military AI onto vulnerable civilian populations.

CHAPTER 3:

FUTURE PERSPECTIVES, OVERSIGHT, AND GLOBAL REGULATORY EFFORTS AND PRACTICAL APPLICATIONS

3.1. Regulatory Efforts under the United Nations Framework and Meaningful Human Control (CCW)

The integration of AI and autonomous technologies into military systems has placed internation-



al security architectures and arms control regimes under unprecedented pressure to adapt. Traditional arms control approaches focus on physically countable and quantifiable conventional or weapons of mass destruction; however, the dual-use nature of AI software blurs the boundaries between military and civilian technologies. This reality has led to deep conceptual and normative divisions in multilateral diplomatic negotiations conducted under the United Nations.

3.1.1. UN CCW GGE Processes and the Stalemate in Legally Binding Treaty Negotiations

The primary multilateral platform for international discussions on Lethal Autonomous Weapon Systems (LAWS) is the Group of Governmental Experts (GGE), established in 2016 under the 1980 Convention on Certain Conventional Weapons (CCW).

Through negotiations within the GGE, the High Contracting Parties endorsed "11 Guiding Principles" in 2019. These principles affirmed that IHL applies fully to autonomous weapons, that human responsibility must be retained throughout the entire life-cycle of a system, and that autonomous weapons should not be anthropomorphized. However, these 11 principles largely reiterate established customary international law and have not been converted into a legally binding treaty.

Given the fundamental uncertainty of autonomous software and the difficulty of establishing individual criminal guilt, robust and objective state responsibility under ARSIWA serves as the indispensable legal mechanism to prevent states from externalizing the catastrophic risks of military AI onto vulnerable civilian populations.

Negotiations within the CCW framework face a structural stalemate. The main drivers of this include:

- **The Consensus Rule:** Adopting a new legally binding protocol under CCW requires the unanimous agreement of all state parties. Leading military powers -such as the US, Russia, Israel, and India- argue that existing IHL rules are sufficient and consistently oppose or veto proposals for a new prohibition treaty.
- **Conceptual and Definitional Ambiguities:** States have failed to reach an agreed technical definition of «autonomy» and «autonomous weapon systems». This dynamic mirrors the challenges seen with the 2021 UN Treaty on the Prohibition of Nuclear



Weapons (TPNW); while many states demand a total ban, the major military powers that actively develop and field these technologies refuse to sign it, creating a divide between legal aspirations and operational realities.

- **Limitations of Article 36 Reviews:** National legal reviews of new weapons conducted under Article 36 of Additional Protocol I are classified as state secrets by most governments. This lack of international transparency prevents Article 36 from acting as an effective substitute for a dedicated regulatory treaty.

3.1.2. Meaningful Human Control (MHC) and the Distributed Military Targeting Cycle (AJP-3.9)

To prevent algorithms from making autonomous life and death decisions, international civil society coalitions and numerous states have advocated for the principle of "Meaningful Human Control" (MHC) as a mandatory regulatory standard. However, detailed legal and operational studies of military targeting show that diplomatic debates often misunderstand MHC by reducing it to a simple, isolated interaction between an operator and a weapon.

Merel Ekelhof (2018, 2019) and Mark Roorda (2015) demonstrate that in modern warfare, control is not exercised solely at the moment of "pulling the trigger", but is distributed across a wider military planning chain through a model known as "Distributed Control".

Under the North Atlantic Treaty Organization's (NATO) Allied Joint Doctrine for Joint Targeting (AJP-3.9), targeting is structured as a six-phase decision-making cycle:

1. **Commander's Objectives, Guidance, and Intent**
2. **Target Development** (Target System Analysis [TSA], vetting, and validation)

3. **Capabilities Analysis** (Weaponeering and Collateral Damage Estimation [CDE])
4. **Commander's Decision, Force Planning, and Assignment**
5. **Mission Planning and Execution** (F2T2EA: Find, Fix, Track, Target, Engage)
6. **Combat Assessment**

- **Target System Analysis (TSA):** The core of targeting takes place in Phase 2 during target development. Determining whether a target is a legitimate military objective (such as a radar site, fuel depot, or command post) requires months of intelligence analysis and TSA evaluations. Consequently, AI integration primarily enters at the level of target discovery and intelligence software, shifting the debate from «autonomous weapons» to «autonomous targeting processes».
- **The Launch Decision (Point of No Return):** Roorda (2015) explains that even if a munition operates autonomously after launch, the commander's initial decision to deploy that system into a specific geographic area with defined target parameters represents the central human choice (the point of no return) that ensures legal compliance.
- **The F-16 Operational Case Study:** Ekelhof (2019) highlights that during night missions in overcast weather, an F-16 fighter jet pilot using GPS-guided bombs does not see the target visually; instead, the pilot releases the weapon when the aircraft enters the pre-calculated weapon envelope upon system prompt. Control does not rest exclusively with the pilot pushing the button, but is distributed across the intelligence analysts, legal advisers, and commanders who validated the target beforehand.

3.1.3. Alternative Regulatory Models: Two-Tier Strategy, Phased Regulation, and Soft Law

To bridge diplomatic impasses and protect civilian populations, three alternative governance models have emerged in international literature:

To prevent algorithms from making autonomous life and death decisions, international civil society coalitions and numerous states have advocated for the principle of "Meaningful Human Control" (MHC) as a mandatory regulatory standard.

1. **The Two-Tier Regulatory Strategy:** Drawing on UN General Assembly Resolution 78/241 and joint appeals by the UN Secretary-General and the President of the International Committee of the Red Cross (ICRC), Erum Iftikhar (2025) outlines a two-tier framework dividing systems into a total ban on unconstrained lethal autonomous machines targeting humans and strict operational limits for semi-autonomous platforms.
2. **Phased and Evidence-Based Regulation:** Nicholas Cromie (2026) proposes a middle path starting with fixed defensive systems in low-risk environments, requiring empirical proof of compliance via black-box audits before expanding to mobile offensive roles.
3. **Soft Law Governance and Multi-Stakeholder Standards:** Gary Marchant (2019) and Gutierrez & Marchant (2021) argue that soft law provides an essential mechanism to address the «Pacing Problem» where technological change outpaces traditional treaty-making through voluntary standards like IEEE P7000.

3.2. Strategic Approaches of Regional and Global Powers

Global powers pursue divergent military AI doctrines to secure operational superiority. While

This institutional tension reflects a broader conflict of interest where ethical aspirations clash with the operational demand for battlefield speed. If the EU relies merely on civilian regulatory declarations without developing sovereign, fully verified defense technologies, it will inevitably be forced during high-tempo crises to deploy opaque, imported "black-box" systems to maintain strategic parity with adversaries.

the US and its allies claim to promote responsible AI governance, China and Russia advance rapid modernization doctrines with minimal treaty constraints, and the European Union (EU) balances civilian ethics with NATO defense interoperability. This section analyzes these strategic postures and the geopolitical competition that incentivizes states to prioritize algorithmic speed over normative consensus.

3.2.1. The US and Allied Doctrines (DoDD 3000.09, RAI, and Defense Partnerships)

The US aims to secure operational and normative leadership by claiming to establish responsible military AI governance standards (the word "claim" has been used since AI-powered weapon systems designed by the US have been subject to controversy). The Department of Defense (DoD) issued Directive 3000.09 (Autonomy in Weapon Systems) in 2012 (updated in 2023), establishing that autonomous and semi-autonomous systems must be designed to allow commanders and operators to exercise "appropriate levels of human judgment over the use of force".

In February 2020, the DoD adopted five Ethical Principles for Artificial Intelligence, implemented through its Responsible AI (RAI) Strategy:

- **Responsible:** Personnel will exercise appropriate levels of judgment and care.
- **Equitable:** Deliberate steps will be taken to minimize unintended bias.
- **Traceable:** Systems possess transparent and auditable methodologies and data.
- **Reliable:** Capabilities have explicit, well-defined uses and safety tests.
- **Governable:** Capabilities can detect errors and disengage or deactivate.

To promote these standards internationally, the US created the AI Partnership for Defense (PfD) with 16 allied countries and developed trustworthiness metrics through the Five Eyes Technical Cooperation Program (TTCP).

Key allies have developed aligned frameworks tailored to their domestic strategic cultures:

- **France:** Formulated the concept of «Controlled AI» (*IA maîtrisée*), emphasizing technological sovereignty, auditable algorithms, and maintaining the commander's direct legal responsibility across all operational distances.
- **Australia:** Developed a «Method for Ethical AI in Defence" based on "Lawful AI", introducing the concept of algorithmic contestability and deploying software tools (such as Athena AI) designed to identify and protect non-targetable sites and persons under IHL.

3.2.2. Doctrines of the People's Republic of China and the Russian Federation

- **China:** China's military modernization focuses on the doctrine of «Intelligentized Warfare" (*zhimeng hua*). While Beijing officially supports a binding international protocol to ban the use of fully autonomous weapons at the UN CCW, its position selectively excludes restrictions on the development, testing, and production of military AI systems, reflecting an asymmetric strategic posture.
- **Russia:** Russian military doctrine prioritizes replacing human soldiers with robotic and

autonomous combat platforms. Moscow opposes binding international bans within the CCW, arguing that existing IHL is sufficient and that restrictive regulations would hinder national defense research and technological modernization.

3.2.3. The EU, NATO, and the Quest for Strategic Autonomy

The EU promotes "Strategic Autonomy" and "Digital Sovereignty" in emerging technologies. However, Simona R. Soare (2023) identifies three structural challenges in the EU's approach:

- **Supply Chain Dependencies:** The EU relies heavily on non-EU suppliers for advanced semiconductors (microelectronics), cloud computing infrastructure, and critical raw materials.
- **Normative Focus:** EU policy focuses primarily on civilian regulatory frameworks (such as the EU AI Act), which creates a normative focus that can limit the development of military operational capabilities.
- **Coordination with NATO:** While NATO advances military AI through its 2021 AI Strategy, the DIANA technology accelerator, and a $1 billion Innovation Fund (NIF) focused on interoperability, overlapping initiatives between the EU and NATO risk duplication across European defense programs.

Ultimately, this institutional tension reflects a broader conflict of interest where ethical aspirations clash with the operational demand for battlefield speed. If the EU relies merely on civilian regulatory declarations without developing sovereign, fully verified defense technologies, it will inevitably be forced during high-tempo crises to deploy opaque, imported "black-box" systems to maintain strategic parity with adversaries. In doing so, the EU risks replicating the severe civilian harms and automation biases seen in modern algorithmic warfare, proving that genuine strategic autonomy cannot rely on passive regulatory rhetoric, but requires verifiable indigenous engineering, strict standard operating procedures, and enforceable human oversight.

3.3. Practical Application and Case Study: Technological Asymmetry and Doctrine Shifts in the Russia-Ukraine War

The Russia-Ukraine War that started in 2014 and intensified in 2022 with Russia's invasion of Ukraine has marked a definitive turning point in modern military history, serving as a large-scale, operational testing ground for algorithmic warfare, AI, uncrewed aerial systems (UAS), and loitering munitions in a high-intensity conflict. Moving beyond asymmetric counter-terrorism applications, the war in Ukraine highlights how low-cost drone innovations and automated targeting systems reshape tactical doctrines, electronic warfare resilience, and the legal limits of automated target acquisition.

3.3.1. Operational Dynamics and the Use of AI-Enabled Platforms in the Russia-Ukraine War

The conflict is characterized by profound technological adaptation, mass proliferation, and the widespread integration of autonomous and semi-autonomous capabilities by both state forces:

The Proliferation of Small Drones and Low-Cost Ammunition: Defying traditional security studies expectations that large, expensive military-grade drones would dominate modern air combat, the war has demonstrated the tactical supremacy of lightweight, man-portable commercial drones and first-person view (FPV) systems. As Kunertova (2023) notes, "drones have boots", providing tactical flexibility down to the platoon level and supplying units with real-time reconnaissance and strike capabilities.

AI-Driven C4ISR and Compressed Decision Loops: Both sides have integrated advanced commercial and military AI platforms for battlefield data fusion and geospatial intelligence. These systems automate target identification, correlate battlefield coordinates, and optimize artillery fire

coordination, dramatically compressing the sensor-to-shooter loop from thirty minutes to mere minutes. (Lee, 2025).

Electronic Warfare (EW) and Counter-Autonomy: The electronic battlespace in Ukraine features intensive GPS spoofing, frequency jamming, and radio-frequency denial. This environment has forced developers to embed higher levels of onboard autonomy and optical machine-navigation algorithms, enabling systems to complete terminal phases without satellite guidance. Furthermore, major military powers like Russia continue to resist external legal constraints, prioritizing unconstrained technological modernization and domestic defense R&D as vital components of their great power identity and strategic parity. (Vallejos et al., 2017; Lee, 2025).

3.3.2. Manufacturer Liability, Ethics, and Standard Operating Procedures for Civilian Protection

The heavy reliance on commercial-off-the-shelf (COTS) components, dual-use artificial intelligence software, and rapid battlefield innovations raises critical legal and regulatory challenges regarding accountability, dual-use supply chains, and adherence to IHL:

Dual-Use Technology Supply Chains and Complicity: Because modern military drones frequently rely on commercial microprocessors, optical sensors, and open-source machine learning libraries (with many components tracing back to global commercial markets), tracking legal responsibility is exceptionally complex. Defense contractors and software developers who knowingly supply dual-use components or unverified targeting algorithms without rigorous Testing, Evaluation, Verification, and Validation (TEVV) face potential scrutiny under international complicity norms and arms transfer regulations. (Lee, 2025).



Ultimately, AI will remain a permanent feature of modern military arsenals, but technological advancement must not dictate the boundaries of international law.

Challenges in Distinguishing Protected Objects and *Hors de Combat* Status: The high speed of algorithmic target recognition in dynamic urban and rural landscapes increases the risk of misclassifying civilian infrastructure (such as agricultural vehicles, energy grids, or non-combatant transport) as military objectives. Furthermore, automated tracking systems struggle to reliably detect surrendering soldiers or wounded personnel (*hors de combat*) when engaged in high-speed terminal maneuvers. (Lee, 2025).



Mandatory Standard Operating Procedures (SOPs) for Accountability: To mitigate civilian harm and prevent violations of distinction and proportionality amidst intense algorithmic warfare, military organizations must enforce strict operational safeguards:

- **Geographic and Temporal Restrictions:** Limiting autonomous search and engagement modes strictly to designated front-line zones where civilian presence is minimized.
- **Pre-Approved Hardware Profiles:** Restricting autonomous terminal attack functions exclusively to verified military assets (such as active radar emitters or heavy armor) while prohibiting autonomous decisions against human personnel.
- **Tamper-Proof Black-Box Logging:** Equipping AI-enabled weapon platforms with secure, unalterable data recorders to log sensor inputs, confidence scores, and algorithmic decisions, ensuring transparency for post-conflict legal audits.
- **Resilient Human Oversight Links:** Maintaining robust, jam-resistant communication channels to guarantee that human commanders retain the ultimate capability to veto, override, or self-destruct autonomous systems prior to impact. (Vallejos et al., 2017; Lee, 2025).

CONCLUSION

The rapid emergence and fielding of military Artificial Intelligence (AI) and autonomous weapon systems have brought modern warfare to a critical turning point, representing a profound shift in decision-making agency that challenges International Humanitarian Law (IHL), criminal accountability, and global security.

At the center of this transformation lies an acute conflict of interest; while geopolitical competition pushes countries to accelerate AI integration for operational speed and precision, states bear binding obligations under IHL to ensure compliance with distinction, proportionality, and precautions. The pursuit of "decision-speed superiority" frequently conflicts with the deliberate, qualitative judgment required on the battlefield.

Recent high-intensity conflicts -such as the widespread deployment of AI-enabled uncrewed systems, automated target recognition, and loitering munitions in the Russia-Ukraine War- illustrate the operational realities and severe risks of algorithmic warfare. The extensive use of autonomous capabilities in contested electronic environments, coupled with AI-driven C4ISR data fusion, highlights the immense pressure on military forces to compress the sensor-to-shooter loop. However, delegating targeting assessments to probabilistic models without robust human oversight risks catastrophic errors, misclassification of civilian infrastructure, and tragic civilian harm.

Legally, software algorithms lack moral agency and *mens rea*, meaning traditional command responsibility cannot apply directly to machines. Accountability remains with human actors, including commanders, software developers, and manufacturers who supply flawed targeting code, alongside states engaging objective responsibility under the UN International Law Commission (ILC) Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA).

To resolve diplomatic stalemates within the UN Convention on Certain Conventional Weapons (CCW), the international community must implement enforceable regulatory frameworks:

1. **A Two-Tier Legal Architecture:** Categorical prohibition of fully autonomous lethal systems targeting humans without meaningful human control, paired with strict operational limits for semi-autonomous platforms.
2. **A Phased Regulatory Approach:** Initial restricted deployment of fixed defensive systems in low-risk environments, expanding only after empirical verification of IHL compliance.
3. **Mandatory Operational Safeguards (SOPs):** Restricting terminal (point of no return) engagement to pre-approved hardware profiles, enforcing strict geographic/temporal corridors, maintaining tamper-proof black-box logging, and ensuring distributed human oversight across military forces' targeting cycle.
4. **Interim Soft Law Standards:** Enforcing multi-stakeholder ethical design principles (e.g., IEEE P7000) through procurement and industrial certifications.

Ultimately, AI will remain a permanent feature of modern military arsenals, but technological advancement must not dictate the boundaries of international law. Only through transparent engineering, rigorous accountability, and strict operational procedures can the principles of distinction, proportionality, and human dignity be protected in 21st century warfare.

REFERENCES

Akkuş, B. (2023). Command accountability for AI weapon systems in the law of armed conflict. *Akademi Sosyal Bilimler Dergisi*, *10*(29), 111–122.

Bode, I. (2024, May 9). Falling under the radar: The problem of algorithmic bias and military applications of AI. *Humanitarian Law & Policy Blog (ICRC)*.

Bode, I., & Huelss, H. (2018). Autonomous weapons systems and changing norms in international relations. *Review of International Studies*, *44*(3), 393–413.

Boutin, B. (2023). State responsibility in relation to military applications of artificial intelligence. *Leiden Journal of International Law*, *36*(1), 133–150.

Cromie, N. (2026). *Lethal autonomous weapons systems and the principles of distinction and proportionality*[Unpublished doctoral dissertation]. The City University of New York (CUNY).

Ekelhof, M. A. C. (2018). Lifting the fog of targeting: 'Autonomous weapons' and human control through the lens of military targeting. *Naval War College Review*, *71*(3), 61–95.

Ekelhof, M. A. C. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy*, *10*(3), 343–348.

Gutierrez, C. I., & Marchant, G. (2021). A global perspective of soft law programs for the governance of artificial intelligence. *SSRN Electronic Journal* (Research Paper No. 3855171), 1–64.

Iftikhar, E. (2025). AI, autonomous weapons, and the crisis in international humanitarian law. *Annals of Human and Social Sciences*, *6*(3), 167–183.

Jacobsen, J. T., & Liebetrau, T. (2023). Artificial intelligence and military superiority: How the 'cyber-AI offensive-defensive arms race' affects the US vision of the fully integrated battlefield. In F. Cristiano, D. Broeders, F. Delerue, F. Douzet, & A. Géry (Eds.), *Artificial Intelligence and International Conflict in Cyberspace* (pp. 135–156). Routledge.

Jevglevskaja, N. (2014–2015). Legal review of new weapons: Origins of Article 36 of AP I. *Finnish Yearbook of International Law*, 25, 109–140.

Kahn, L., Horowitz, M. C., & Samotin, L. (2026). *What is human in judgment? Comparing automation bias and algorithm aversion between the United States Military Academy and the general public* [Working Paper].

Kahn, L. (2022). Lethal autonomous weapon systems and respect for human dignity. *Frontiers in Big Data*, 5, 999293.

Kunertova, D. (2023). Drones have boots: Learning from Russia's war in Ukraine. *Contemporary Security Policy*, *44*(4), 576–591.

Laiq, M. F. (2026). Artificial intelligence in future warfare: Ethical frameworks and the regulation of lethal autonomous weapons. *IEEE Transactions on Technology and Society*, *7*(1), 1–25.

Lee, J. (2025). Autonomous weapons, war crimes, and accountability. *North Carolina Journal of International Law*, *49*(2), Article 4.

Marchant, G. (2019). 'Soft law' governance of artificial intelligence. *AI PULSE Papers*, UCLA eScholarship, 1–20.

Milanovic, M. (2025/2026). The use of artificial intelligence in military operations: Complicity, state responsibility and individual criminal responsibility under international law. *SSRN Electronic Journal* (Research Paper No. 6183440), 1–23.

Moy, W. R., & Gradon, K. T. (2023). Artificial intelligence in hybrid and information warfare: A double-edged sword. In F. Cristiano, D. Broeders, F. Delerue, F. Douzet, & A. Géry (Eds.), *Artificial Intelligence and International Conflict in Cyberspace* (pp. 47–74). Routledge.

Oimann, A.-K., & Salatino, A. (2025). Command responsibility in military AI contexts: Balancing theory and practicality. *AI and Ethics*, *5*(4), 1757–1767.

Pacholska, M. (2023). Military artificial intelligence and the principle of distinction: A state responsibility perspective. *Israel Law Review*, *56*(1), 3–23.

Roorda, M. (2015). NATO's targeting process: Ensuring human control over (and lawful use of) 'autonomous' weapons (*ACIL Research Paper No. 2015-06*). Amsterdam Center for International Law, 152–168.

Soare, S. R. (2023). Algorithmic power? The role of artificial intelligence in European strategic autonomy. In F. Cristiano, D. Broeders, F. Delerue, F. Douzet, & A. Géry (Eds.), *Artificial Intelligence and International Conflict in Cyberspace* (pp. 77–108). Routledge.

Spadaro, A. (2023). A weapon is no subordinate: Autonomous weapon systems and the scope of superior responsibility. *Journal of International Criminal Justice*, *21*(5), 1119–1136.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Tsybulenko, E., & Kajander, A. (2022). Customary international humanitarian law and Article 36 of Additional Protocol I to the Geneva Conventions: A stopgap regulator of autonomous weapons systems? *TalTech Journal of European Studies*, *12*(2), 87–112.

Winter, E. (2020). The compatibility of the use of autonomous weapons with the principle of precaution in the law of armed conflict. *The Military Law and the Law of War Review*, *58*(2), 240–273.

Winter, E. (2022). The compatibility of autonomous weapons with the principles of international humanitarian law. *Journal of Conflict & Security Law*, *27*(1), 1–20.

Zając, M. (2023). AWS compliance with the ethical principle of proportionality: Three possible solutions. *Ethics and Information Technology*, *25*(1), Article 13.