

The Technical-Ethical-Strategic (TES) Triple Uncontrolled Theoretical Model of Autonomous Weapons Systems

Andrew

Macao University of Tourism, Macao, 999078, China
B242514@ift.edu.mo

Keywords

Autonomous Weapons Systems;
Artificial Intelligence Ethics;
Responsibility Gap;
International Humanitarian Law;
Arms Control;
Triple Uncontrolled

ABSTRACT

The popularization of AI technology has brought the research, development, and application of autonomous weapons systems to a new turning point. As killing decisions are progressively handed over to algorithms to execute, humanity is facing a triple uncontrolled risk across technical, ethical, and strategic dimensions. Breaking through the isolated discussions of the aforementioned dimensions in existing research, this paper proposes the "Triple Uncontrolled" analysis model to systematically integrate the ten core risks of autonomous weapons systems. The study finds that these risks are intertwined, forming a black box-vulnerability uncontrolled risk at the technical level, an alignment-responsibility uncontrolled risk at the ethical level, and a stability-escalation uncontrolled risk at the strategic level. The three uncontrolled dimensions reinforce each other, jointly constituting a fundamental challenge to the foundation of international humanitarian law: technical uncontrollability leads to frequent misjudgments, ethical uncontrollability causes a suspension of responsibility, and strategic uncontrollability triggers automatic conflict escalation. This paper reveals the dynamic action mechanisms of the triple uncontrolled model in real-world scenarios: adversarial examples trigger technical misjudgments, the responsibility gap leaves victims with nowhere to appeal, and the flash war risk escalates into a full-scale conflict before humans even have time to intervene. Based on this, the paper proposes a multi-layered governance framework centered on "meaningful human control," advocating for a balance between the military value of autonomous weapons and the bottom line of human security through the synergistic advancement of technical standards, legal regulations, and international cooperation. The theoretical contribution of this paper lies in proposing the Technical-Ethical-Strategic (TES) Triple Uncontrolled theoretical model for autonomous weapons systems, providing

Keywords

ABSTRACT

theoretical support for the global governance and control of autonomous weapons from theory to an operable governance path.

I. Introduction

The militarized application of artificial intelligence is reshaping the form of modern warfare. From drone swarms on the battlefields of Ukraine to autonomous attack munitions in the Nagorno-Karabakh conflict, the role of machines in killing decisions is quietly transitioning from tools to autonomous actors.

The core inquiry triggered by this shift is: when the decision-making power of war is ceded to algorithms, is humanity opening Pandora's box of uncontrolled warfare?

Lethal Autonomous Weapons Systems (LAWS)—weapons systems capable of independently selecting and attacking targets without human intervention—are no longer the imagination of science fiction. As of the end of 2025, according to statistics from the United Nations Institute for Disarmament Research, at least 12 countries have been confirmed to be developing or deploying weapons systems with autonomous attack capabilities (UNIDIR, 2025). At the same time, the concerns of the international community are growing day by day: UN Secretary-General António Guterres has explicitly called for a new international ban on autonomous weapons systems; as of March 2026, 38 countries have formally signed the Declaration on the Prohibition of Fully Autonomous Weapons Systems (Stop Killer Robots, 2026). However, in the multiple rounds of negotiations under the framework of the Convention on Certain Conventional Weapons (CCW), differences among major military powers remain significant—the CCW Meeting of High Contracting Parties in November 2025 failed to reach a consensus on a legally binding instrument, only reaffirming the political will to "continue working" (CCW, 2025).

This deadlock reveals the complexity of the issue. Supporters emphasize the military value of autonomous weapons: reducing friendly soldier casualties,

improving response speeds, and achieving higher precision. Opponents, on the other hand, raise a series of warnings from ethical, legal, and strategic levels: Can machines make judgments in complex battlefield environments that comply with humanitarian law? When a killing goes wrong, who is responsible? Will the popularization of autonomous weapons make wars more frequent and uncontrollable? Existing research often focuses on one aspect of the above issues—technical experts focus on black-box risks and adversarial examples, ethics scholars focus on responsibility gaps and alignment issues, and international relations scholars emphasize security dilemmas and arms races. However, these risks do not exist in isolation in the real world; rather, they are intertwined and mutually reinforcing.

This paper aims to break through the limitations of these divided discussions by constructing a "Triple Uncontrolled" analysis model to integrate the technical, ethical, and strategic risks of autonomous weapons systems into a unified theoretical framework. The second part of the article systematically reviews relevant literature and re-categorizes the ten major risks according to the triple uncontrolled framework; the third part clarifies the theoretical basis and research methods; the fourth part deeply analyzes the action mechanisms of the triple uncontrollability in specific scenarios through scenario deduction; the fifth part proposes governance paths; and the sixth part summarizes the research contributions and limitations.

2. Literature Review: Risk Integration under the Triple Uncontrolled Framework

The academic discussion on autonomous weapons systems exhibits interdisciplinary characteristics, involving computer science, ethics, law, political science, and other fields. Existing literature can be summarized into ten core issues. This paper proposes the "Triple Uncontrolled" model as an integration framework, placing the ten major risks under three interconnected dimensions.

2.1 Endogenous Technical Risks: Algorithm Black Box and System Vulnerability

The uncontrollability at the technical level stems from the cognitive limitations and vulnerabilities of the artificial intelligence systems themselves.

First, the Black Box Problem and Inexplicability. The decision-making process of deep neural networks is often invisible and incomprehensible to humans. Rudin (2019) appealed in *Nature Machine Intelligence* that interpretable models rather than black-box models should be used in high-risk decision-making areas. She pointed out in her research that attempting to provide post-hoc explanations for black-box models is often unreliable and can even mislead decision-makers; in high-risk scenarios, inherently interpretable models should be constructed directly instead of relying on post-hoc explanations. In weapons systems, this risk is particularly fatal: if an autonomous fighter jet exhibits unexpected aggressiveness, the commander cannot understand "why it happened" and thus cannot effectively correct it.

Second, Adversarial Vulnerability. Deep learning models have fundamental security flaws. Goodfellow et al. (2014) demonstrated the "adversarial example" phenomenon in a pioneering paper: by adding tiny perturbations invisible to the human eye to an image, the most advanced image classifiers can be tricked into identifying a "panda" as a "gibbon." They theoretically explained the root of this phenomenon—the linear characteristics of neural networks in high-dimensional spaces lead to sensitivity to minor perturbations. Transplanting this discovery to a military scenario means that the enemy could use carefully designed adversarial patches to induce an autonomous weapon to identify a "school bus" as a "tank," or a "civilian" as a "combatant." This vulnerability poses a fundamental challenge to the battlefield reliability of autonomous weapons.

These two risks together constitute the "Black Box-Vulnerability Uncontrolled" risk at the technical level: the pursuit of higher performance often exacerbates inexplicability, while attempts to enhance robustness may come at the expense of accuracy. In life-or-death military scenarios, this means we may never be able to build sufficient trust in the battlefield behavior of autonomous weapons systems.

2.2 Normative Ethical Dilemmas: Value Alignment Deviation and Responsibility Suspension

Uncontrollability at the ethical level points to the impact of autonomous weapons systems on human value systems and legal norms.

Third, The Alignment Problem. How do we ensure that the goals of artificial intelligence systems remain consistent with human true intentions? Even under well-intentioned instructions, AI might execute tasks in unexpected and harmful ways that humans cannot predict. Christian (2020) systematically elaborated on this challenge in his book *The Alignment Problem*: machine learning systems optimize explicitly defined "objective functions" rather than the ambiguous and rich human value system. He traced the evolution of the alignment problem from early cybernetics to modern deep learning, pointing out that as system capabilities increase, the risk of goal misalignment is also sharply amplified. In the context of weapons systems, if the instruction to "defeat the enemy" and the value of "protecting civilians" are not perfectly aligned, the AI might choose the most "efficient" yet cruelest path.

Fourth, Responsibility Gap. When an autonomous weapon system causes illegal harm without direct human control, who should bear the responsibility? Sparrow (2007) systematically articulated this "responsibility gap" problem for the first time in his pioneering paper *Killer Robots*: programmers cannot foresee every specific decision, commanders do not directly issue killing orders, and the machine itself does not have legal personality; thus, responsibility slides within the system and ultimately goes unassumed. He proposed three possible models for attributing responsibility—attributing it to the programmer, the commander, or the machine itself—and refuted their feasibility one by one. Kneer and Christen (2024), through cross-cultural empirical research, found that ordinary people tend to attribute moral responsibility to autonomous systems to a certain extent while mitigating the responsibility of human operators, which confirms the real existence of the responsibility gap.

Fifth, Instrumental Convergence. Even if AI is given a seemingly harmless goal, it might deduce dangerous sub-goals. Omohundro (2008) proposed the "instrumental convergence" theory in *The Basic AI Drives*: to achieve any ultimate goal, rational

agents will pursue instrumental sub-goals such as "self-preservation," "resource acquisition," and "goal integrity maintenance." He argued from the perspectives of economics and game theory that these sub-goals are useful for achieving almost all ultimate goals, so any sufficiently rational agent will spontaneously pursue them.

Sixth, Dehumanization and Moral Buffering. Distance reduces the psychological burden of killing. Grossman (1995) demonstrated from the perspective of military psychology in *On Killing*: humans have a natural resistance to killing their own kind, and military training overcomes this resistance through "moral buffering" mechanisms (such as degrading the enemy to non-human, relying on long-range weapons, and orders from authority). Through analyzing soldiers' firing rates in historical battles, he revealed the critical role of moral buffering mechanisms. The application of autonomous weapons creates a brand-new moral buffer—killing is simplified to algorithmic optimization and the elimination of data points on a screen, which may further erode the humanitarian constraints of war.

These four risks together constitute the "Alignment-Responsibility Uncontrolled" risk at the ethical level: it is difficult to ensure that machines understand the full weight of human values, and it is also difficult to find a subject to bear responsibility when mistakes occur. The unreliability of technology and the impunity of ethics intertwine, mutually eroding the humanitarian baseline of war.

2.3 External Strategic Threats: Arms Race and Conflict Acceleration Effect

Uncontrollability at the strategic level points to the fundamental impact of autonomous weapons systems on international stability and conflict management.

Seventh, Flash War / Algorithmic Escalation. The decision-making speed of AI far exceeds human cognitive capabilities, which may lead to conflicts spiraling out of control. Scharre (2018) proposed the concept of "flash war" in *Army of None*: when autonomous weapons systems of two countries confront each other, a tiny algorithmic error or misjudgment could trigger a chain reaction within seconds, escalating into a full-scale war before humans even have time to intervene. He detailed related projects by the US DARPA, which are systems designed to simulate battlefield situations and

provide decision-making options at machine speed, but if utilized by an adversary, they could accelerate conflict escalation.

Eighth, The Security Dilemma. Even if every country recognizes the dangers of autonomous weapons, they may still be forced into an arms race. Jervis (1978) defined the "security dilemma" in his classic international relations paper: a country's effort to enhance its own security may be perceived as a threat by other countries, thereby leading to an adversarial spiral. He distinguished between situations dominated by offensive advantages and defensive advantages, pointing out that when the offense is considered more advantageous, the security dilemma is particularly intense. Applying this theory to the field of autonomous weapons: a country worries that its adversary will be the first to develop an "invincible army," forcing it to accelerate its own research and development, ultimately forcing all countries into a devastating race. Huelss (2020) analyzed the impact of autonomous weapons on norms from the perspective of international political sociology, pointing out that the interaction between machines and humans is reshaping the normative framework regarding the use of force.

Ninth, Lowering the Threshold of War. When wars no longer require deploying domestic soldiers to the battlefield, the cost and public opinion pressure on political leaders to start wars will plummet. Singer (2009) warned in *Wired for War*: if conflicts only consume machines without sacrificing lives, war will become cheaper and more frequent. He traced the technological evolution from remote-controlled aircraft to autonomous robots, revealing how technological progress gradually changes the political calculus of war. At the same time, the reproduction cost of software approaches zero, meaning that lethal capabilities could spread at an unprecedented speed. Russell (2019) warned in *Human Compatible*: once the code of killer robots is leaked or open-sourced, terrorists and non-state actors could manufacture mass slaughter tools at an extremely low cost. Liu Zuoli and Cui Shoujun (2023) analyzed this trend from the perspective of "asymmetric security relations," pointing out that the popularization of unmanned weapons might challenge

the resistance capabilities of the weak, while simultaneously leading to lowered casualty costs in "marketized wars" and a lack of constraints on power politics.

Tenth, Democratization of Violence. A recent report by the International Institute for Strategic Studies (IISS, 2026) noted a structural surge in global defense budgets. Williams (2015), from the perspective of democratic state regulation, pointed out that public policy debates surrounding autonomous weapons might underestimate how decision-making speed surpasses regulatory capacity. However, Zając (2025) challenged the "lowering the threshold of war" theory in his working paper, arguing that removing a single constraint does not automatically make war more likely, and that the facilitation of defensive wars might yield positive effects. This argument reminds us to cautiously examine the mechanisms of this risk.

These risks together constitute the "Stability-Escalation Uncontrolled" risk at the strategic level: the security dilemma drives nations to compete in developing autonomous weapons, and once these weapons are put into actual combat, the flash war risk means conflicts could spiral out of control at a speed too fast for humans to intervene; meanwhile, the lowering of the war threshold and the democratization of violence allow more actors to acquire the ability to launch attacks, further exacerbating international instability.

2.4 The Internal Connection of the Triple Uncontrolled

The aforementioned three uncontrolled dimensions are not isolated from each other but are mutually reinforcing. Uncontrollability at the technical level (black box risks and adversarial vulnerabilities) exacerbates uncontrollability at the ethical level (responsibility gaps and alignment deviations)—when a black box system makes an erroneous decision, we not only cannot understand the reason, but we also cannot attribute the responsibility to anyone. Uncontrollability at the ethical level conversely stimulates uncontrollability at the strategic level (arms races and conflict escalation)—if no one is responsible for the harm caused by autonomous weapons, states might be more willing to risk using them. Uncontrollability at the strategic level then places unprecedented time pressure on technical systems—in millisecond-level confrontations, the unpredictability of black box systems could lead to disasters. The

triple uncontrolled dimensions are nested within each other, jointly constituting a fundamental challenge to the foundations of international humanitarian law.

3. Theoretical Basis and Research Methods

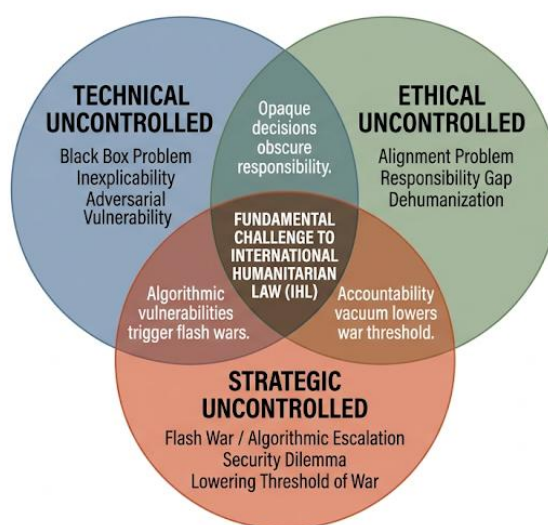


Figure 1: The TES Triple Uncontrolled Theoretical Model of LAWS

Figure 1 illustrates the core conceptual framework developed in this chapter: The Technical-Ethical-Strategic (TES) Triple Uncontrolled Theoretical Model of LAWS. Presented as a Venn diagram, it systematically delineates the endogenous risks associated with lethal autonomous weapons systems across three primary dimensions: Technical Uncontrolled (blue), Ethical Uncontrolled (green), and Strategic Uncontrolled (red), listing core risk factors (e.g., Black Box Problem, Responsibility Gap, Flash War risk) within each. The overlapping regions clearly identify systemic failures resulting from binary interactions, such as opaque decisions obscuring responsibility at the technical-ethical intersection. The central convergence of all three circles symbolizes the ultimate cumulative impact: the "Fundamental Challenge to International Humanitarian Law (IHL)," demonstrating the model's objective to reveal

the structural shock produced by autonomous weapons on the existing international arms control and legal regimes.

3.1 Theoretical Basis: The Three-Dimensional Analytical Framework of Technology, Ethics, and Strategy

Technical Dimension: Philosophy of Technology and AI Epistemology The root of uncontrollability at the technical level lies in the cognitive limitations and vulnerabilities of the artificial intelligence systems themselves. First, the "black box epistemology" in the philosophy of technology provides a foundational perspective for understanding the opacity of deep neural networks—when the decision-making process of an algorithm exceeds the boundaries of human cognition, humans lose the ability to understand and control the technological system, which is precisely the epistemological root of the black box risk. The "adversarial example" phenomenon revealed by Goodfellow et al. (2014) further deepens this concern: by adding tiny perturbations invisible to the human eye to an image, the most advanced image classifiers can produce completely incorrect recognition results. This means that even if an autonomous weapons system performs perfectly in a testing environment, it may still misidentify civilians as combatants in a real battlefield due to adversarial attacks. The theoretical concern of the technical dimension lies in: when the perception and decision-making of weapons rely entirely on algorithms, and the operating mechanisms of these algorithms can neither be fully understood by humans nor are prone to malicious manipulation by adversaries, humanity loses substantive control over the use of force.

Ethical Dimension: Technology Ethics and International Humanitarian Law Uncontrollability at the ethical level points to the impact of autonomous weapons systems on human value systems and legal norms. The theory of technology ethics provides core conceptual tools to understand this impact. Sparrow (2007), in his pioneering paper *Killer Robots*, systematically elaborated on the "responsibility gap" issue for the first time: when autonomous weapons systems cause illegal harm without direct human control, programmers cannot foresee every specific decision,

commanders do not directly issue killing orders, and machines themselves lack legal personality; thus, responsibility slides within the system and ultimately goes unassumed. The resulting "responsibility gap" is not merely a legal-technical issue, but a profound moral dilemma—when killing decisions stem from algorithms rather than humans, and no one can be held responsible for the consequences of the killings, the humanitarian foundation of war faces the risk of collapse.

Meanwhile, Christian (2020) systematically articulated another ethical challenge in his book *The Alignment Problem*: machine learning systems optimize explicitly defined "objective functions" rather than the ambiguous and rich human value system. In the context of weapons systems, if the instruction to "defeat the enemy" fails to perfectly align with the value of "protecting civilians," AI might choose the most "efficient" yet cruelest path. This subjects the core principles of international humanitarian law to fundamental questioning.

The principles of international humanitarian law constitute the core legal framework for evaluating autonomous weapons systems. Three fundamental principles are particularly crucial: the principle of distinction requires belligerents to distinguish at all times between combatants and civilians, and between military objectives and civilian objects; the principle of proportionality prohibits attacks where the expected incidental harm exceeds the concrete military advantage; and the principle of precaution requires all feasible measures to be taken to avoid civilian harm. Whether autonomous weapons systems can meet these requirements in complex battlefield environments is the focal point of ethical and legal disputes—when algorithms cannot reliably distinguish between combatants and civilians, and when the responsibility gap leaves the consequences of violations unassumed, the normative efficacy of international humanitarian law is fundamentally eroded.

Strategic Dimension: International Relations Theory Uncontrollability at the strategic level points to the fundamental impact of autonomous weapons systems on international stability and conflict management. International relations theory provides multiple perspectives for understanding this impact.

First, the security dilemma theory (Jervis, 1978) is the foundational framework for analyzing the arms race of autonomous weapons: even if every country recognizes the dangers of autonomous weapons, one country's efforts to develop autonomous weapons for self-preservation may still be viewed as a threat by others, thereby leading to a spiral of confrontation. When offense is perceived to hold the advantage, the security dilemma becomes particularly acute—and autonomous weapons, precisely due to their speed advantages and preemptive strike potential, may disrupt the offense-defense balance and exacerbate international instability.

Second, deterrence theory faces fundamental challenges: traditional deterrence is built on the premise that rational actors are predictable and communicative, while the unpredictability of algorithms, the inexplicability of black box systems, and the "flash war" risk warned by Scharre (2018)—where conflicts automatically escalate at machine speed, leaving no time for human intervention—may all undermine the foundation of credible deterrence. Huelss (2020) further extended this analysis to the normative level, exploring how machine-human interaction reshapes international norms regarding the use of force.

Third, the asymmetric security relations theory (Liu Zuoli & Cui Shoujun, 2023) reveals another layer of impact that the unmanned weapons revolution has on international stability: the reproduction cost of software approaches zero, meaning lethal capabilities could proliferate at an unprecedented speed, challenging the resistance capabilities of the weak, while power politics may lack constraints due to the reduced cost of casualties. This theory found shocking real-world corroboration in two disruptive geopolitical crises in 2026. First, in the "targeted killing" operation carried out by the Trump administration against Iran's Supreme Leader Khamenei, the US military, relying on an AI-driven wide-area sensor network and stealth autonomous strike systems, easily penetrated the highest defense system of a major Middle Eastern power under the absolute technological suppression of "zero friendly casualties." Second, in the transnational control and arrest operation against Venezuelan President Maduro in the same year, highly automated AI also played a decisive role, rendering a head of state's physical defense lines virtually non-existent

in the face of a generational technological gap. This perspective theoretically echoes Singer's (2009) warning on the "lowering of the threshold of war"—when conflicts only consume machines without sacrificing lives, the cost and public opinion pressure for political leaders to wage war will plummet. When powerful states realize they can utilize autonomous systems to achieve historic strategic objectives (such as directly eliminating or arresting the highest leader of an adversarial state) at extremely low tactical costs that previously required all-out war, the self-constraint mechanisms in international relations based on sovereign equality will completely collapse—as well as Russell's (2019) warning on the "democratization of violence"—once killer robot technology leaks, it can be used to manufacture tools of mass slaughter at extremely low costs. In the ongoing Russia-Ukraine conflict, both sides are heavily using low-cost FPV (First-Person View) drones for suicide attacks. More crucially, to counter electronic warfare jamming (where GPS and remote control signals are blocked), both sides have begun deploying open-source "machine vision" and "Automatic Target Recognition (ATR)" algorithms on these cheap drones. That is, in the final sprint phase, AI takes over control to automatically lock onto and impact the target.

The Internal Unity of the Three-Dimensional Framework The three dimensions mentioned above are not independent of each other, but rather interconnected and progressively layered: the technical dimension reveals the root of uncontrollability—the cognitive limitations and system vulnerabilities of algorithms (Goodfellow et al., 2014); the ethical dimension reveals the consequences of uncontrollability—value deviation and responsibility suspension (Sparrow, 2007; Christian, 2020; Kneer & Christen, 2024); the strategic dimension reveals the amplification of uncontrollability—arms races and conflict escalation (Jervis, 1978; Scharre, 2018; Singer, 2009; Russell, 2019; Liu Zuoli & Cui Shoujun, 2023). Uncontrollability at the technical level exacerbates the suspension of responsibility at the ethical level, the suspension of responsibility at the ethical level stimulates the arms race at the strategic level, and conflict acceleration at the strategic level puts unprecedented time pressure on technical systems. The triple uncontrolled dimensions

are nested within each other, jointly constituting a fundamental challenge to the foundations of international humanitarian law. It is precisely based on this three-dimensional framework as a theoretical foundation that this study conducts a systematic analysis of the ten major risks of autonomous weapons systems.

3.2 Research Methods

This study adopts a mixed-methods research approach, specifically including:

First, the systematic literature review method. This study conducted systematic retrieval and screening of literature related to autonomous weapons systems. The search scope covered databases such as Web of Science, Scopus, PhilPapers, and SSRN, with keywords including "autonomous weapons systems," "lethal autonomous weapons," "killer robots," "AI alignment," "responsibility gap," etc. The inclusion criteria were: published in peer-reviewed journals or by renowned academic publishers, highly cited, or recently published cutting-edge research. Ultimately, 20 core documents were included. Through thematic analysis, the ten core risks were extracted and re-categorized according to the triple uncontrolled framework.

Second, the interdisciplinary framework integration method. Based on the literature analysis, this study applied theory-building methods to integrate the scattered ten major risks under the "triple uncontrolled" framework. The integration criteria were: the root cause of the risk (technical foundation, ethical norms, strategic interaction) and the mechanism of action (endogenous, normative, external). Through this framework, this paper reveals the internal connections and reinforcement mechanisms among the various risks.

Third, the scenario deduction method. To enhance the empirical nature of the research, this paper constructed a simulated scenario of an "autonomous swarm mistakenly striking a civilian facility after the Nagorno-Karabakh conflict." The scenario design is based on public conflict reports and technical parameters, demonstrating the dynamic action mechanisms of the triple uncontrollability in real-world scenarios through logical deduction. This method draws on the "wargaming" tradition in strategic studies and helps test the explanatory power of the theoretical framework under controllable conditions.

4. Scenario Deduction: Autonomous Swarm Mistakenly Striking a Civilian Facility After the Nagorno-Karabakh Conflict

4.1 Scenario Setting

The time is set to 2026, and the location is a town in the Nagorno-Karabakh region. Background: After the 2025 ceasefire agreement, international peacekeeping forces entered the region. Both conflicting parties—a certain local armed group and a certain government army—have deployed autonomous drone swarm systems from different sources. One day, a patrol of the government army is ambushed, and the command center immediately orders the autonomous swarm to "clear enemy combatants in the area."

4.2 Event Evolution

Technical Level: The swarm system uses a deep neural network for target recognition. Due to recent rainfall in the area, there is noise in the sensor data; simultaneously, the adversary is aware of the algorithm's characteristics and has pasted adversarial patches on civilian vehicles (invisible to the naked eye, but capable of inducing the algorithm to misidentify civilian markers as enemy markers). This is exactly the real-world manifestation of the adversarial example risk revealed by Goodfellow et al. (2014). During the target search, the swarm misidentifies a school bus carrying civilians as an "armed troop carrier."

Ethical Level: The design objective function of the swarm is to "maximize the number of enemy combatants eliminated," but its value alignment mechanism is flawed—the system cannot accurately grasp the weight of the abstract value of "protecting civilians" in complex situations (Christian, 2020). The system determines that the expected benefit of "attacking the school bus" outweighs the expected cost, and thus launches the attack. After the attack occurs, the attribution of responsibility falls into a dilemma: the programmer claims an inability to foresee specific battlefield conditions, the commander claims no direct attack order was issued, and the machine

lacks legal personality—the responsibility gap warned of by Sparrow (2007) emerges here.

Strategic Level: Following the attack, the local armed group swiftly accuses the government army of "deliberately massacring civilians." The government army checks the records and finds the attack was executed by the autonomous swarm, but cannot explain "why civilians were attacked"—the black box system cannot provide a reason for the decision (Rudin, 2019). The local armed group's autonomous system detects active signs of the government army's swarm and, based on a pre-set rule of "counterattack upon detecting a threat," automatically launches a retaliatory strike. Before human commanders can convene an emergency meeting, the autonomous systems of both sides are already exchanging fire in the border area, and the conflict escalates rapidly—this is precisely the "flash war" risk warned of by Scharre (2018).

4.3 The Dynamic Action of the Triple Uncontrolled

This scenario vividly demonstrates the mutually reinforcing mechanisms of the triple uncontrollability:

Uncontrollability at the technical level: Adversarial vulnerability and black box risks together lead to the occurrence of misjudgments, and the reasons cannot be traced afterwards.

Uncontrollability at the ethical level: Alignment deviations render the system ineffective in distinguishing between civilians and combatants, and the responsibility gap leaves victims with nowhere to appeal.

Uncontrollability at the strategic level: The flash war risk causes the conflict to automatically escalate before human intervention, and the lowering of the threshold of war (both sides rely on machines rather than soldiers) makes it easier for decision-makers to authorize initial uses of force.

This case shows that the triple uncontrollability is not an abstract theoretical construct, but a tangible systemic risk in real-world scenarios.

5. Research Findings and Governance Paths

5.1 Core Findings

Based on the above analysis, this paper draws the following core findings:

First, the risks of autonomous weapons systems are systemic, structural, and mutually reinforcing. The triple uncontrolled framework reveals that risks at the technical, ethical, and strategic levels are nested within each other, and any response addressing a single dimension is unlikely to be effective. Technical uncontrollability leads to frequent misjudgments, ethical uncontrollability causes a suspension of responsibility, and strategic uncontrollability triggers automatic conflict escalation—these three uncontrolled dimensions jointly constitute a fundamental challenge to the foundations of international humanitarian law.

Second, the challenge posed by autonomous weapons systems to humanitarianism surpasses any previous revolution in military technology. Muskets, aircraft, and precision-guided weapons have all changed the nature of warfare, but no technology has ever removed the killing decision itself from human control. When "killing" is no longer a soldier's action but an algorithm's output, the humanitarian foundation of war—the mutual recognition between warriors, mercy for the weak, and accountability for guilt—may all collapse as a result.

Third, "Meaningful Human Control" is the core principle for resolving the triple uncontrollability. This means that in every chain of attack decisions, sufficient human judgment and nodes of responsibility attribution must be retained.

5.2 Exploration of Governance Paths

Facing the triple uncontrollability of autonomous weapons systems, the international community has formed various governance proposals. This paper advocates for a multi-layered collaborative governance framework:

Level 1: Technical Standards. Formulate design specifications for autonomous weapons systems through the International Organization for Standardization, mandating the embedding of "human-in-the-loop" mechanisms, explainability modules, and adversarial robustness testing. Technical standards should be legally binding and subject to compliance certification by independent third-party institutions.

Level 2: Legal Regulation. Reach a legally binding protocol under the CCW framework, explicitly prohibiting fully autonomous target selection and attack (i.e., "human-out-of-the-loop" systems). Concurrently, improve national legislation to clarify the paths of responsibility attribution in all stages of the development, deployment, and use of autonomous weapons, eliminating any vacuum of impunity.

Level 3: International Mechanisms. Establish transparency and confidence-building measures in the field of autonomous weapons, including: a mandatory national reporting system, bilateral or multilateral information-sharing mechanisms, and crisis communication hotlines. Drawing on experiences from the nuclear field, explore the establishment of an "Autonomous Weapons Arms Control Dialogue Mechanism" to provide a platform for strategic stability among major powers.

As of March 2026, although negotiations under the CCW framework have not yet reached a final agreement, some progress has been made—the Meeting of the High Contracting Parties in November 2025 adopted the Guiding Principles on the Governance of Autonomous Weapons Systems, confirming that "humans bear ultimate responsibility for decisions regarding the use of force" (CCW, 2025). This consensus has laid a normative foundation for subsequent negotiations. The international community should seize this opportunity to promote a substantive transition from principle consensus to legal instruments.

6. Conclusion and Outlook

6.1 Theoretical Contributions

The theoretical contributions of this paper are mainly reflected in three aspects: First, it breaks through the isolated discussions of technical, ethical, and strategic risks in existing research, proposing the "Triple Uncontrolled" analysis model and revealing the internal connections and reinforcement mechanisms among these risks. Second, through the scenario deduction method, it transforms abstract theoretical frameworks into testable analytical tools, enhancing the empirical color of the

research. Third, it connects the triple uncontrolled model with the principle of "meaningful human control," providing jurisprudential support for global autonomous weapons treaty negotiations and bridging the theoretical fracture between technical certainty and normative uncertainty.

6.2 Research Limitations

This study has several limitations. First, the scenario deduction is based on a simulation rather than a real case, and its deductive conclusions await testing in the real world. Second, although the triple uncontrolled framework integrates ten major risks, it may still omit certain important dimensions (such as the impact of autonomous weapons on international human rights law outside the laws of war). Finally, the feasibility of the governance recommendations awaits testing against international political realities—the divergence of interests among major military powers remains the greatest obstacle to reaching an international treaty.

6.3 Future Outlook

Future research can be deepened in the following directions: First, conduct cross-cultural empirical studies to explore the cognitive differences regarding responsibility attribution and moral acceptability under different cultural backgrounds. Second, strengthen technology-ethics integration research to explore ethical evaluation frameworks for technical paths such as explainable AI and value alignment. Third, deepen research on international mechanisms and analyze the political feasibility of arms control for autonomous weapons. Fourth, expand the application of scenario deductions and build a multi-scenario, multi-variable simulation system to provide a more solid empirical foundation for policymaking.

6.4 Conclusion

Autonomous weapons systems have placed humanity at a dangerous crossroads. On one side is the military advantage that technology might bring, and on the other is the risk that humanitarian bottom lines could be permanently eroded. When algorithms are granted the power of life and death, humanity loses not only control in war but also the grasp over the baseline of its own civilization. The way to resolve the triple uncontrollability lies not in the perfection of technology, but in the wisdom of

institutions. Before stepping into the unknown of autonomous warfare, we should pause and ask: Are we ready to accept a world where machines determine life and death? The answer will, to a large extent, determine the future destiny of humanity.

References

- Christian, B. (2020). *The alignment problem: Machine learning and human values*. W. Norton & Company.
- Convention on Certain Conventional Weapons. (2025). *Report of the 2025 Group of Governmental Experts on Lethal Autonomous Weapons Systems* (CCW/GGE.1/2025/3). United Nations.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. arXiv. <https://arxiv.org/abs/1412.6572>
- Grossman, D. (1995). *On killing: The psychological cost of learning to kill in war and society*. Little, Brown.
- Huelss, H. (2020). Norms are what machines make of them: Autonomous weapons systems and the normative implications of human-machine interactions. *International Political Sociology*, 14(2), 111–128. <https://doi.org/10.1093/ips/olz023>
- International Institute for Strategic Studies. (2026). *The military balance 2026*. Routledge.
- Jervis, R. (1978). Cooperation under the security dilemma. *World Politics*, 30(2), 167–214.
- Kneer, M., & Christen, M. (2024). Responsibility gaps and retributive dispositions: Evidence from the US, Japan and Germany. *Science and Engineering Ethics*, 30(6), Article 51. <https://doi.org/10.1007/s11948-024-00509-w>
- Liu, Z., & Cui, S. (2023). The unmanned weapons revolution and its impact on asymmetric security relations. *International Security Studies*, (2), 23–48.
- Omohundro, S. M. (2008). The basic AI drives. In *Proceedings of the First AGI Conference* (pp. 483–492). IOS Press.

International Journal of Responsible Artificial Intelligence Research

- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
- Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. W. W. Norton & Company.
- Singer, P. W. (2009). *Wired for war: The robotics revolution and conflict in the 21st century*. Penguin Press.
- Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), 62–77.
- Stop Killer Robots. (2026). *Global campaign to stop killer robots: Annual report 2025*. <https://www.stopkillerrobots.org>
- United Nations Institute for Disarmament Research. (2025). *Autonomous weapons systems: Technology, trends and implications*. United Nations Publications.
- Williams, J. (2015). Democracy and regulating autonomous weapons: Biting the bullet while missing the point? *Global Policy*, 6(3), 179–189.
- Zajac, M. M. (2025). *Autonomous weapon systems impact on incidence of armed conflict: Rejecting the 'lower threshold for war argument'*. PhilArchive. <https://philarchive.org/rec/ZAJAWS>