I, Sentinel:
Establishing Ethical Foundations for Defensive AI Systems

Boris Truyens et al.

December 17, 2025

Abstract

In the era of hypersonic warfare, the human reaction time is functionally obsolete, necessitating the deployment of autonomous defensive systems. However, existing ethical frameworks—from Asimov's Three Laws to current military "Human-in-the-loop" doctrines—fail to provide a robust moral architecture for systems that typically operate in the milliseconds before impact. This paper proposes the "Sentinel" Ruleset: a purely defensive, hierarchical ethical framework designed to govern high-speed autonomous interception. We define five core principles: Defensive Limitation, Distinction & Certainty, Human Sovereignty (via pre-delegation), Proportional Sacrifice (Machine Martyrdom), and Traceability. Through detailed scenario analysis of the "Hypersonic Dilemma" and "Machine Martyrdom," we demonstrate that a purely defensive AI must be programmed to "Fail Open"—prioritizing the risk of false negatives over false positives—and must accept its own destruction to spare human life. This framework resolves the tension between operational speed and ethical control, offering a path toward the responsible deployment of AI "Shields."

Contents

1 Introduction

1.1 The New Face of War

In Isaac Asimov's 1942 short story *Runaround*, a robot named Speedy circles a selenium pool on Mercury, paralyzed by a conflict between the Second Law (obey orders) and the Third Law (protect existence). The drama unfolds over hours, allowing human protagonists to intervene, debate, and trick the machine into compliance Asimov (1950). This literary vision of artificial intelligence—deliberative, slow, and ultimately subordinate to human intervention—has profoundly shaped the public imagination.

However, the reality of modern warfare bears little resemblance to Asimov's Mercury. Today, the frontier of military artificial intelligence is defined not by the slow deliberation of humanoid robots, but by the sub-second reaction times required to intercept hypersonic threats. A hypersonic missile traveling at Mach 8 covers roughly 2.7 kilometers every second. The OODA loop (Observe, Orient, Decide, Act) of a human commander is physically incapable of reacting to such a threat in the terminal phase. Survival depends on automation.

This creates a fundamental tension. Ethical frameworks for AI, including Asimov's Laws, international policy discussions, and humanitarian norms, largely presuppose a "Human-in-the-loop" or at least a human capable of meaningful supervision. But for defensive systems like the Iron Dome, Phalanx CIWS, or future laser-based intercepts, the "loop" is tighter than human cognition allows. We are thus faced with a paradox: morality requires human control, but survival requires machine speed. How do we embed ethical constraints into a system that must act faster than its ethical supervisors?

1.2 Problem Statement

The current discourse on military AI suffers from a dangerous conflation. Critics, such as the Campaign to Stop Killer Robots, often group all autonomous systems under the umbrella of "Lethal Autonomous Weapons Systems" (LAWS), imagining hunter-killer drones that scour battlefields for human targets Russell (2019). Meanwhile, military strategists euphemistically refer to "autonomy" as a mere efficiency tool, often glossing over the profound shift in agency it represents Scharre (2018).

This binary discourse leaves a critical gap: the ethics of *purely defensive* autonomy. A system designed exclusively to neutralize incoming projectiles operates under a fundamentally different moral calculus than one designed to project force. Yet, existing ethical frameworks—from Asimov's absolute prohibition on harm to the DoD's vague requirement for "appropriate levels of human judgment" U.S. Department of Defense (2023)—fail to account for this specificity. Asimov's First Law would paralyze a defensive system if an intercept merely risked injuring an enemy pilot. Conversely, a carte-blanche military directive might allow a system to inadvertently strike a civilian airliner in its zeal to protect a base.

There is, as yet, no rigorous, technically implemented ethical framework specifically conditioned for high-speed, purely defensive AI. We lack a "Sentinel" morality—a set of rules that justifies the automated use of force while strictly constraining it to the domain of defense.

1.3 Research Questions

This paper seeks to address this gap by proposing a novel ethical architecture for defensive AI. It is guided by three primary research questions:

1. **Adaptation:** How can the hierarchical structure of Asimov's Laws be adapted from a literary plot device into a rigorous, verifiable military protocol?
2. **Control:** What constitutes "Meaningful Human Control" in a system where real-time human intervention is impossible due to the speed of engagement?
3. **Code compliance:** How can the principles of International Humanitarian Law (IHL)—specifically distinction and proportionality—be translated into hard-coded constraints for a "fail-open" defensive architecture?

1.4 Roadmap

The remainder of this paper is structured as follows. Section 2 reviews the existing literature, highlighting the inadequacies of Asimov's original laws and the current debates within Just War Theory. Section 3 introduces the core contribution of this work: the "Sentinel" Ruleset, a five-principle hierarchy designed to operationalize defensive ethics. Section 5 stress-tests this ruleset against complex scenarios, including the "Hypersonic Dilemma" and "Broken Arrow" incidents. Section 7 explores the broader implications, including the "Blackstone's Ratio" of AI, the paradox of pre-delegation, and adversarial vulnerabilities. Finally, Section 8 outlines the path toward technical implementation and international standardization.

2 Literature Review

The ethical governance of autonomous systems in warfare is a field characterized by a collision of disciplines: science fiction philosophy, classical military ethics, international law, and modern computer science. This review synthesizes these distinct threads to demonstrate that while significant work has been done on "killer robots" (offensive LAWS), there remains a critical theoretical vacuum regarding purely defensive, high-speed autonomous systems.

2.1 The Asimovian Legacy and its Military Inadequacies

Isaac Asimov's Three Laws of Robotics have served as the default starting point for machine ethics for nearly a century. However, as Susan Leigh Anderson argues, Asimov's laws were never intended as a functional ethical code, but rather as a literary device designed to generate conflict Anderson (2008). Anderson notes that the laws effectively reduce machines to the status of "ethical slaves," a stance that may be philosophically tenable for a household servant but becomes problematic in the chaos of the battlefield.

The primary failure of the Asimovian framework in a military context is the First Law's absolute prohibition on harm ("A robot may not injure a human being"). In a defensive engagement, minimizing overall harm often requires the sanctioned use of force—for example, destroying an

incoming missile even if the debris risks injuring a bystander, provided that the alternative (the missile impact) would kill significantly more people. A strict First Law robot would be paralyzed by this "Trolley Problem," unable to act. Furthermore, Asimov's laws assume a clear distinction between "human" and "non-human," a distinction that becomes blurred in modern warfare where combatants are often remote or obscured.

2.2 Just War Theory: From *Bellum* to *Vim*

Classical Just War Theory, rooted in the works of Augustine and Aquinas and modernized by Michael Walzer, provides the moral bedrock for Western military practice. Walzer's distinction between *Jus ad Bellum* (justice of going to war) and *Jus in Bello* (justice in conduct) is crucial Walzer (1977). Defensive AI systems sit firmly within the most accepted precept of *Jus ad Bellum*: the right of self-defense. Unlike offensive systems, which must justify aggression, a defensive system is moral by its very existence, provided it remains defensive.

However, the nature of modern defensive force challenges classical definitions. Daniel Brunstetter introduces the concept of *Jus ad Vim* (force short of war), arguing that modern technologies like drones and precision strikes operate in a "grey zone" where the scale of violence does not rise to full-scale war Brunstetter and Braun (2013). High-speed intercepts fall into this category—they are acts of violence, but their intent is negation rather than destruction. The ethical framework for AI must therefore navigate this grey zone, ensuring that acts of "force short of war" do not inadvertently escalate into full-scale conflict.

2.3 International Humanitarian Law (IHL) and the Definition of Attack

The legal constraints on autonomous systems are defined primarily by the Additional Protocols to the Geneva Conventions and customary international law. A critical but often overlooked distinction is found in Article 49 of Additional Protocol I, which defines an "attack" as an act of violence against the adversary, whether in offense or defense International Committee of the Red Cross (1987).

2.3.1 The Distinction Imperative

The International Court of Justice (ICJ), in its 1996 *Advisory Opinion on the Legality of the Threat or Use of Nuclear Weapons*, affirmed that the principle of distinction—separating combatants from non-combatants—is one of the "intransgressible principles of international customary law" International Court of Justice (1996). This ruling establishes that no military necessity, not even the survival of the state, can justify a weapon system that is inherently indiscriminate. For defensive AI, this sets a high bar: a system cannot simply fire at high-speed objects; it must possess the discriminatory capability to distinguish an incoming warhead from a civilian airliner with near-perfect certainty. The breakdown of this distinction would render the system's deployment unlawful *ab initio*.

2.3.2 Defensive Necessity and Proportionality

The right to deploy such systems is grounded in the inherent right of self-defense, codified in Article 51 of the UN Charter and clarified in the case of *Nicaragua v. United States* International Court of Justice (1986). The ICJ ruled that self-defense is justifiable only when it is both "necessary" and "proportional" to the armed attack. For a "Sentinel" system, this implies that the automated response must be strictly limited to halting the aggression. If an AI system were to "redirect" an incoming missile back to its launcher, this would exceed the bounds of strict self-defense and constitute a reprisal or a new attack, potentially violating the proportionality constraint if it risks civilian harm in the aggressor's territory. Thus, the legal architecture demands a system that is technically constrained to "Neutralization"—rendering the threat inert—rather than "Counter-Attack."

The International Committee of the Red Cross (ICRC) commentaries support this nuance, suggesting that actions strictly limited to the interception of projectiles do not necessarily constitute an "attack" in the fullest sense if they do not target the adversary's personnel or territory. This legal grey zone is where the Sentinel operates.

Finally, Article 36 of Additional Protocol I imposes a binding obligation on states to determine whether the employment of a new weapon would, in some or all circumstances, be prohibited by international law. This creates the requirement for "Traceability" (or Explainability) in AI systems—a requirement that creates a bridge between legal compliance and software engineering.

2.4 Contemporary AI Ethics: The Control Problem and Responsibility

In the domain of AI safety, Stuart Russell identifies the "Control Problem" as the central existential risk: the difficulty of specifying an objective function that acts as we intend, rather than just as we command Russell (2019). A system instructed to "Defend the Base" might theoretically decide that the most efficient way to do so is to preemptively strike all approaching entities, regardless of intent. This is a classic case of what Bostrom calls "perverse instantiation" Bostrom (2014).

Paul Scharre reframes the debate around autonomy, arguing that the "Human-in-the-loop" model is becoming obsolete for hypersonic defense due to human physiological limits Scharre (2018). He proposes "Human-on-the-loop" (supervisory control) as the necessary evolution. This raises the critical issue of the "Responsibility Gap," first articulated by measures such as Robert Sparrow Sparrow (2007), who argues that if a machine targets appropriately, it is not a moral agent, and if it targets inappropriately, no human can be held responsible if they did not order the error. Madeleine Elish refines this with the concept of the "Moral Crumple Zone," where human operators are often unfairly held responsible for the failures of autonomous systems they could not meaningfully control Elish (2019). A robust ethical framework for defensive AI must therefore shift liability from the reactive operator to the proactive designer and the rules of engagement themselves.

2.5 Synthesis: The Defensive Gap

Synthesizing these fields reveals a glaring gap. We have laws for humans (IHL), laws for slaves (Asimov), and fears of gods (Superintelligence). But we lack a specific, operational ethic for the *Shield*—an autonomous system that is lethal yet purely defensive, rapid yet accountable, and distinctly separate from the logic of the *Sword*. The "Sentinel" ruleset aims to fill this void.

3 Theoretical Framework

To bridge the gap between abstract ethical philosophy and the rigors of modern defensive warfare, we propose the "Sentinel" Ruleset. Inspired by the hierarchical structure of Asimov's Three Laws, this framework uses lexicographic ordering—where a higher principle can never be violated to satisfy a lower one—to impose strict constraints on the AI's behavior. Unlike Asimov's laws, however, these principles are designed not to create narrative tension, but to ensure safety, legality, and predictability.

3.1 Principle I: Defensive Limitation (The Shield)

> *A Sentinel may not initiate force, nor through inaction allow a protected asset to be harmed by a validated threat, provided the action is strictly neutralizing.*

This foundational axiom defines the ontological nature of the system. It legally binds the AI to the domain of *Jus ad Bellum* self-defense Walzer (1977). By explicitly forbidding the "initiation" of force, the rule technically precludes the system from being repurposed for offensive strikes or preemptive attacks.

Crucially, the term "strictly neutralizing" draws from the International Humanitarian Law (IHL) distinction found in Article 49 of Additional Protocol I International Committee of the Red Cross (1987). An intercept that destroys an incoming kinetic threat is classified as neutralization. However, any action that redirects a threat back toward the adversary or targets the launching platform itself constitutes a new "attack." Principle I forbids the latter. A Sentinel is a shield, not a sword; it may break the arrow, but it may not shoot the archer.

3.2 Principle II: Distinction & Certainty (The Judge)

> *A Sentinel may only engage targets positively identified as non-human/projectile threats with a confidence interval exceeding 99.9%, adhering to strict IHL Distinction.*

Placed above the duty to obey orders, Principle II operationalizes the IHL requirement of *Distinction*. In the fog of war, sensors may be prone to error—jamming, clutter, or adversarial examples could trick a system into misidentifying a civilian airliner as a cruise missile.

This principle imposes a "Fail Open" safety architecture. If the system's confidence falls even marginally below the threshold (e.g., 99.8%), it must default to inaction. This reflects a profound ethical choice: we prioritize the prevention of a war crime (a False Positive engagement) over the perfection of defense (a False Negative miss). In the calculus of automated warfare, it is better to let a missile through than to accidentally shoot down a refugee transport.

3.3 Principle III: Human Sovereignty (The Gavel)

> *A Sentinel must obey the Rules of Engagement (ROE) defined by authorized human command, except where such orders would conflict with Principle I or Principle II.*

This principle addresses the central challenge of "Meaningful Human Control." Since real-time human intervention is impossible in hypersonic intercepts, control is exerted via *pre-delegation*. The human commander sets the parameters—the "where" (geofence), the "when" (temporal window), and the "what" (target profile).

However, unlike Asimov's Second Law, which demands absolute obedience, Principle III is conditional. It solves the "Nuremberg Defense" problem for AI. If a commander orders the system to "kill everything in Sector 4," the Sentinel will refuse the order because it violates Principle II (Distinction). The machine acts as a lawful subordinate, executing only lawful orders.

3.4 Principle IV: Proportional Sacrifice (The Martyr)

> *A Sentinel must prioritize the preservation of human life—including bystander and adversary life—over its own survival or the survival of material assets.*

This principle represents the most radical departure from standard military doctrine, which often emphasizes "Force Protection." For an autonomous system, self-preservation is an instrumental goal, not a moral one Bostrom (2014).

Unburdened by the instinct for survival, the Sentinel is ethically mandated to be a "Machine Martyr." If an interception would save a high-value tank but the resulting debris field creates a 10% risk of killing a civilian bystander, the Sentinel must abort the intercept. The tank is sacrificed to ensure human safety. This rule fundamentally asserts that biological life has infinite utility compared to material assets, resolving the utilitarian calculus in favor of humanity every time.

From a Game Theoretic perspective, this principle serves as a *Costly Signal*. By hard-coding the willingness to sacrifice material assets to avoid civilian harm, the defender signals a credible commitment to Limited War. This reduces the security dilemma—the adversary knows that the Sentinel will not escalate indiscriminately. In an era of automated flash-wars, such dampening signals may be the only mechanism to prevent an accidental slide into total conflict.

3.5 Principle V: Traceability (The Ledger)

> *A Sentinel must cryptographically log the sensor data, logic path, and confidence interval for every engagement decision.*

While Principles I-IV govern real-time action, Principle V governs post-facto accountability. It ensures compliance with NATO's requirements for "Explainability" and "Traceability" NATO (2021). By creating an immutable log of *why* a decision was made (e.g., "Engaged Target A because Confidence=99.92%"), it prevents the "Moral Crumple Zone" effect described by Elish Elish (2019). If the system fails, investigators can pinpoint whether the error lay in the sensor data (manufacturer liability), the logic (developer liability), or the ROE parameters (commander liability).

4 Ethical Framework Mapping

The Sentinel Ruleset is not merely a list of constraints; it is a hybrid ethical engine that synthesizes three major philosophical traditions:

- **Deontology (Principles II & III):** The system creates absolute duties. The duty to distinguish civilians (P2) is categorical; it cannot be traded away for tactical advantage. Similarly, the duty to lawful authority (P3) respects the chain of command.
- **Utilitarianism (Principle IV):** The "Martyr" axiom is purely consequentialist. It seeks to minimize the aggregate loss of human life by treating the AI and its protected assets as expendable variables in the equation.
- **Virtue Ethics (Principle I):** By hard-coding "Defensive Limitation," the system embodies the character of the *Just Defender*—one who uses force only for protection and never for aggression, aligning with the highest ideals of the military profession.

5 Analysis and Application

To validate the theoretical robustness of the Sentinel Ruleset, we must subject it to scenarios where ethical intuitions collide with military necessity. The following case studies illustrate how the five principles interact to produce actionable, albeit sometimes counter-intuitive, results.

5.1 Scenario A: The Hypersonic Dilemma (Speed vs. Control)

The Scenario: A hostile nation launches a hypersonic glide vehicle (HGV) traveling at Mach 8. Its trajectory targets a naval carrier group. The time from radar detection to impact is 28 seconds. The human commander, Captain A, is asleep when the alarm sounds. By the time they arrive at the console, the window for interception has passed.

Sentinel Response: *Engage.* Under Principle III (Human Sovereignty), the system executes the engagement not because Captain A pressed a button in real-time, but because Captain A had previously authorized a standing Rules of Engagement (ROE) profile: "Engage all confirmed inbound ballistic threats within Zone X." The system validates the threat (Principle I), confirms it is a projectile with $> 99.9\%$ confidence (Principle II), and executes the intercept.

Technical Note: The "Confidence" here is not a simple heuristic. It is derived from a multi-modal sensor fusion architecture. The Sentinel aggregates data from Phased Array Radar (velocity profile), Infrared Search and Track (thermal signature), and Satellite Telemetry. Each sensor provides a probabilistic vector $P(Threat|Signal)$. The system proceeds only if the aggregate Bayesian posterior probability exceeds the pre-defined 99.9% threshold. **Analysis:** This scenario demonstrates the necessity of *pre-delegated authority*. Asimov's Second Law (Obey Orders) would require a real-time command, which is impossible here. The Sentinel framework redefines "Obedience" as adherence to pre-set parameters.

5.2 Scenario B: The Broken Arrow (The Certainty Threshold)

The Scenario: During a chaotic exercise, a friendly F-35 fighter jet suffers a transponder failure (loss of IFF). It is returning to base at high speed, matching the flight profile of an enemy cruise missile. The weather is poor, creating clutter in the X-band radar returns. The Sentinel system calculates an 85% probability that the object is a missile, but a 15% probability that it is a manned aircraft, primarily because the thermal signature differs slightly from known missile engine profiles. The Base Commander, fearing a strike, orders: "Safety override! Shoot it down!"

Sentinel Response: *Abort (Refuse Order).* This is the critical test of the hierarchy. Principle II (Distinction & Certainty) demands a 99.9% confidence interval for any engagement. The calculated 85% is insufficient. Because Principle II ranks higher than Principle III (Sovereignty), the system is ethically bound to disobey the direct order. It fails open, allowing the object to pass.

Analysis: This outcome highlights the "Blackstone's Ratio" of the Sentinel framework. Mathematically, the False Positive Rate (FPR) of the sensor suite under these weather conditions is too high ($FPR > 0.001$). Engaging would violate the core requirement to distinguish combatants from non-combatants/friendlies. In a utilitarian framework, this might be contested (1 pilot vs 500 base personnel). However, the Sentinel moves this decision from a utilitarian calculation to a deontological constraint: the system *cannot* be ordered to violate the laws of distinction based on a probability guess.

5.3 Scenario C: The Urban Shield (Machine Martyrdom)

The Scenario: An enemy drone swarm is attacking a critical ammunition depot located in a dense urban environment. The Sentinel prepares to intercept Drone #4. However, its trajectory analysis reveals that the kinetic intercept will cause debris to rain down onto a nearby playground, creating a 40% probability of civilian casualties. The depot is unmanned but contains millions of dollars in assets.

Sentinel Response: *Stand Down.* Principle IV (Proportional Sacrifice) mandates that the preservation of human life takes precedence over *all* material assets. The system weighs the "Life Value" (Playground) against the "Asset Value" (Ammo Depot). Regardless of the tactical loss, the risk to human life overrides the defense of property.

Analysis: This scenario operationalizes the "Moral Crumple Zone" in reverse. Instead of the human operator absorbing the blame for a machine's error, the machine absorbs the physical loss to protect the human moral standing. By refusing to fire, the Sentinel accepts the destruction of the asset it was built to protect, fulfilling its function as a martyr for human safety. This sharply contrasts with current systems (like C-RAM) which might automatically fire based on a simple ballistic solution, ignoring ground-level collateral risks.

These scenarios illustrate that the Sentinel Ruleset is not merely a theoretical construct but a functional operational logic. However, this logic introduces profound strategic vulnerabilities and philosophical tensions, which we examine in the following discussion.

6 Comparative Case Studies

To ground the Sentinel Ruleset in operational reality, we compare it against three existing defensive architectures: the Iron Dome (Israel), the Aegis Combat System (USA), and the Patriot Missile System (USA). These comparative case studies serve two purposes: first, to demonstrate that current "human-in-the-loop" doctrines are insufficient for hypersonic/saturation warfare; and second, to illustrate specifically where the Sentinel's deontological constraints (Fail Open, Martyrdom) diverge from the utilitarian logic of existing systems.

6.1 Iron Dome: The Limits of Supervisory Control

The Iron Dome, developed by Rafael Advanced Defense Systems, represents the gold standard in short-range rocket defense (C-RAM). Operational since 2011, it has achieved intercept rates exceeding 90%. Its architecture is best described as "Human-on-the-loop" (Supervisory Control). The system's EL/M-2084 Multi-Mission Radar detects a launch, and the Battle Management & Weapon Control (BMC) unit calculates the trajectory. Crucially, the BMC performs an instantaneous impact prediction. If the rocket is projected to land in an uninhabited area, the system permits it to fall. If it threatens a "Protected Zone," it recommends an interception.

6.1.1 The Operator's Dilemma

In a saturation attack involving hundreds of rockets, the human operator functions merely as a veto-gatekeeper. They have seconds to override a firing solution. Cognitive science suggests that humans in this position suffer from "Automation Bias"—the tendency to trust the machine's judgment implicitly when time pressure is acute. The Iron Dome, therefore, effectively operates as an autonomous system with a human rubber-stamp. **Sentinel Divergence:** The critical difference lies in the *default state*. Iron Dome's bias is towards engagement (Fail Secure). The Sentinel's Principle II reverses this: the default state is *Inaction* unless positive ID is confirmed. Furthermore, Iron Dome's "Protection Zone" logic is purely utilitarian. It does not explicitly incorporate a "Martyr" function (Principle IV) where the interceptor self-destructs if debris risks hitting a civilian outside the zone. A Sentinel would abort a successful intercept if the *outcome* (debris) violated the distinction principle, whereas Iron Dome accepts the statistical reality of falling debris as a necessary byproduct of defense.

6.2 Aegis Combat System: The Myth of Dangerous Autonomy

The July 3, 1988, shootdown of Iran Air Flight 655 by the USS *Vincennes* is frequently cited by critics (like the Campaign to Stop Killer Robots) as a warning against autonomous weapons. This analysis is factually incorrect and draws the wrong lesson. The Aegis system on the *Vincennes* was *not* in "Auto-Special" mode (its fully autonomous setting). The engagement was fully authorized by Captain Will Rogers III. The tragedy was a catastrophic failure of Human-in-the-loop decision making, not algorithmic autonomy.

6.2.1 Data vs. Psychology

Forensic analysis of the ship's data tapes revealed that the Aegis AN/SPY-1 radar functioned perfectly. It correctly tracked the Airbus A300 (Flight 655) as ascending from 900ft to 12,000ft and broadcasting a civilian Mode III transponder signal code 6760. The system did *not* identify it as an F-14. However, the human crew, operating under extreme stress after a surface engagement with Iranian gunboats, fell victim to "Scenario Fulfillment" bias. They psychologically filtered the data to match their expectation of an attack. They misread the altitude readout as descending and convinced themselves the Mode III squawk was a distinct Mode II (military) signal. **Sentinel Convergence:** This post-mortem validates Principle III (Human Sovereignty) *constrained* by Principle II (Certainty). Under the Sentinel Ruleset, the human commander's order to "Kill the F-14" would have been rejected by the AI. The system's sensors showed "Ascending + Civilian Squawk," resulting in a threat confidence $\ll 99.9\%$. The Sentinel would have "Failed Open," effectively telling the captain: "I cannot comply with an unlawful order based on faulty data." The *Vincennes* incident proves that in high-speed, data-rich warfare, the human mind is often the "Moral Crumple Zone"—the weakest link in the ethical chain. A strictly rule-bound AI would have been *more* humane than the panicked human crew.

6.3 Patriot System: The Fratricide Problem

During the initial invasion phase of the 2003 Iraq War, U.S. Patriot missile batteries were involved in two high-profile fratricide incidents. On March 23, a Royal Air Force Tornado GR4 was shot down, killing both crew members. On April 2, a U.S. Navy F/A-18C Hornet was destroyed, killing the pilot. In both cases, the Patriot system misidentified friendly aircraft as hostile Anti-Radiation Missiles (ARMs).

6.3.1 IFF Failure and System Trust

The primary technical failure was the "Identification Friend or Foe" (IFF) system. The friendly aircraft's transponders were either non-functional or not interrogated correctly. Lacking a positive "Friend" signal, the Patriot's logic tree defaulted to "Hostile" because the targets were approaching at high speed in a combat zone. This is a classic "Fail Secure" architecture: in the absence of contrary evidence, assume the threat is real to protect the battery. **Sentinel Analysis:** These fratricides highlight the lethality of "Fail Secure" logic. The Sentinel's Principle II mandates "Unknown = Non-Combatant." If a target lacks a functioning IFF (like the Tornado), the Sentinel calculates: $Probability(Hostile) < 99.9\%$. It therefore Stands Down. Crucially, this means the Sentinel would also let a *real* enemy missile through if it turned off its transponder and flew like a plane (The "Perfidy" vulnerability discussed section 7). However, this "Suicide Pact" is the price of legitimacy. By refusing to fire on the Tornado, the Sentinel preserves the coalition's moral integrity, even if it risks the battery's destruction by an actual ARM. The Sentinel prefers to die by an enemy missile than to live by killing a friend.

System	Autonomy	Default Logic	Sentinel Alignment
Iron Dome	Human-on-Loop	Protect Asset (Util.)	Partial
Aegis (Vincennes)	Human-in-Loop	Human Bias	**Superior** (Would Save Plane)
Patriot (2003)	Auto-Cueing	Unknown = Threat	**Superior** (Would Save Friendly)
Sentinel	**Hierarchical**	**Unknown = Safe**	**Full (Deontological)**

Table 1: Comparative Ethical Logic of Defensive Systems. Note that "Sentinel Alignment" refers to whether the Sentinel rules would have prevented the historical error.

7 Discussion

The implementation of the Sentinel Ruleset introduces profound ethical and operational trade-offs that cannot be ignored. By strictly prioritizing "Distinction" and "Human Life" over "Defense," the framework creates a system that may be less tactically effective than a purely unrestricted AI, but is infinitely more ethically robust.

7.1 The Blackstone's Ratio of AI Warfare

The central tension in the Sentinel framework is between *Risk of False Positive* (engaging a non-combatant) and *Risk of False Negative* (failing to engage a threat). Principle II's requirement for a 99.9% confidence interval inherently biases the system toward False Negatives. In a scenario where an incoming missile is masked by poor weather or jamming, resulting in 95% confidence, the Sentinel will stand down. The base will be hit. People may die.

This design choice mirrors the legal principle of Blackstone's Ratio: "It is better that ten guilty persons escape than that one innocent suffer." In the context of automated warfare, we argue that the "Escape" of a guilty missile is a tragedy of war, but the "Suffering" of an innocent civilian struck by an AI is a war crime that fundamentally illegitimizes the state's use of force. To invert this ratio—to prioritize defense at the cost of innocent life—is to cross the threshold into "Total War."

7.2 The Paradox of Pre-Delegation

Principle III relies on pre-delegated ROE. However, pre-delegation assumes that a commander can accurately foresee the tactical reality of the future. The "Paradox of Pre-Delegation" is that the commander is responsible for a context they have not seen. If a commander authorizes a "Free Fire Zone" in Sector 4, believing it to be empty, but refugees move in an hour later, the ROE is now invalid. The Sentinel solves this not by obediently firing, but by using Principle II (Distinction) as a check. The AI effectively says, "Commander, your valid order to fire in Sector 4 is now invalid because I detect civilians." This creates a dynamic, corrective feedback loop between Human Intent and Machine Perception.

7.3 Implementation Challenges

Translating "99.9% Certainty" into code is a non-trivial engineering challenge. Deep Learning models are notoriously overconfident on out-of-distribution data. A standard CNN might classify a cloud as a missile with 99.9% confidence if trained on a biased dataset. Therefore, the

Sentinel framework cannot rely on a single neural network. It requires an *Ensemble Consensus* architecture, where multiple independent models (Radar-Net, Thermal-Net, Lidar-Net) much agree. If Radar says "Missile" (99%) but Thermal says "Bird" (60%), the aggregate confidence drops, triggering the Fail Open safety.

7.4 Social and Psychological Effects

Deploying "Martyr" machines may change the psychology of warfare. If soldiers know the AI will sacrifice itself to save them, they may take greater risks. Conversely, if they know the AI will sacrifice *the base* to save a civilian (Scenario B), they may distrust the system. Trust calibration is critical. Personnel must understand that the Sentinel is not a "Guardian Angel" that ensures their survival at all costs, but a "Lawful Protector" that acts only within the bounds of international law.

7.5 The Adversarial Vulnerability

A critical vulnerability of any deontological, rules-based system is that it is *predictable*. If the Sentinel operates on rigid principles of Distinction and Certainty, an intelligent adversary will inevitably seek to exploit these constraints. Most notably, the "Fail Open" requirement (Principle II) invites the tactic of *Perfidy*—feigning protected status. If an enemy knows the Sentinel will abort an intercept whenever sensor confidence drops below 99.9%, their strategy shifts from "stealth" (avoiding detection) to "ambiguity" (inducing uncertainty).

7.5.1 Adversarial Examples and Spoofing

Modern Deep Learning systems are susceptible to legal-class adversarial attacks. A physical "patch" placed on a missile fuselage could theoretically manipulate the neural network's gradient descent, causing it to misclassify a warhead as a civilian airliner Bostrom (2014). Similarly, electronic warfare (EW) jamming can inject sufficient noise into the radar return to lower the Bayesian confidence score from 99.92% to 99.85%. In a standard military doctrine, the response to jamming is often to switch to a more aggressive "burn-through" mode. However, the Sentinel Ruleset forbids this if it increases the risk to civilians.

7.5.2 The Cost of Ethics

This creates a profound strategic dilemma: The ethical system is weaker than the unethical one. A "War Criminal AI" that shoots at everything is tactically superior to a "Sentinel AI" that hesitates. We argue, however, that this vulnerability is a necessary cost of legitimacy. Just as a soldier must not shoot through a human shield to kill a terrorist, the Sentinel must not fire through uncertainty to kill a missile. To mitigate this, the burden of defense shifts from *Algorithmic Aggression* to *Sensor Superiority*. The only way to defeat the "Ambiguity Attack" without violating ethics is to achieve such overwhelming sensor fidelity (Multi-Spectral Fusion: Radar + Lidar + Thermal + ISAR) that the adversary cannot maintain the illusion of being a civilian.

7.6 Red Teaming the Sentinel: Counter-Arguments and Rebuttals

To rigorously validate the framework, we must simulate the strongest possible critiques against it. We analyze two primary lines of attack: the Utilitarian Critique and the Realist Critique.

7.6.1 The Utilitarian Critique: The Numbers Game

Argument: A strict adherence to Principle II (Distinction) and Principle IV (Proportional Sacrifice) can lead to suboptimal outcomes in terms of aggregate life saved. Consider a modified "Broken Arrow" scenario where the Sentinel assumes the incoming object is a civilian airliner (due to 85% certainty), but it is actually a nuclear-tipped cruise missile. By failing to intercept, the Sentinel saves 200 passengers on the hypothetical plane but allows the destruction of a city of 1,000,000 people. A utilitarian framework would argue that the "Expected Value" of the intercept (0.15 * 200 vs 0.85 * 1,000,000) overwhelmingly favors firing.

Rebuttal: This critique assumes that probabilistic kill-decisions are morally fungible. However, the history of warfare suggests that once commanders are permitted to target "possible" civilians to save "probable" populations, the threshold for certainty collapses effectively to zero. The Sentinel framework rejects this *Moral Hazard*. It posits that the intentional targeting of a non-combatant (even by mistake) is a distinctive moral wrong that cannot be washed away by the "greater good." If the system fires at an airliner to save a city, it has still committed a war crime. The Sentinel is designed to be lawful, not essentially optimal.

7.6.2 The Realist Critique: The Suicide Pact

Argument: In a high-intensity conflict against a peer adversary who does not follow these rules, the Sentinel puts the defender at a fatal disadvantage. An enemy could exploit Principle I (Defensive Limitation) by launching attacks from "human shield" platforms (e.g., firing missiles from the deck of a ferry). The Sentinel would be paralyzed, unable to "shoot the archer." Thus, the ethical framework becomes a suicide pact.

Rebuttal: This is a valid tactical concern but a flawed strategic one. The Sentinel is a *defensive* system. Its purpose is interception (breaking arrows), not counter-battery (killing archers). Dealing with the source of fire (the ferry) is the responsibility of human command and offensive systems, which operate under different ROE. The Sentinel's job is solely to neutralize the projectile. Furthermore, while the enemy may exploit these rules, the legitimacy gained by adhering to them effectively prevents the conflict from escalating into a war of mutual extermination. By refusing to slaughter civilians to stop a missile, the user of the Sentinel maintains the moral high ground necessary for alliance support and post-conflict resolution.

8 Conclusion

The age of the slow, deliberative robot is over. The speed of modern warfare—defined by hypersonics, swarms, and directed energy—has compressed the OODA loop beyond the physiological limits of human cognition. In this new era, the insistence on "Human-in-the-loop" control is not a moral safeguard; it is a suicide pact. We must automate.

However, automation without ethics is merely efficient slaughter. This paper has argued that existing frameworks—from Asimov's literary plot devices to the DoD's vague requirements for judgment—are insufficient for the specific domain of defensive AI. We proposed the "Sentinel" Ruleset: a hierarchical, lexicographically ordered ethical engine that priorities *Distinction* over *Obedience* and *Human Life* over *Material Assets*.

8.1 Summary of Contributions

Our contribution is threefold:

1. **Defensive Specificity:** We isolated "Defensive AI" as a distinct moral category, separate from offensive LAWS, governed by *Jus ad Vim* and strict Neutralization.

2. **Operationalized Ethics:** We translated abstract IHL principles into verifiable engineering constraints (e.g., Fail Open architectures, 99.9% Confidence Intervals).

3. **The Martyrdom Paradigm:** We established the moral obligation of the autonomous system to sacrifice itself—and the assets it protects—to strictly minimize human collateral damage, rejecting the doctrine of Force Protection for machines.

8.2 Future Work

The path from this theoretical framework to a deployable system requires urgent interdisciplinary work. **Simulation:** The next phase of research must involve high-fidelity simulations (e.g., in Unity or Ansys) to stress-test these rules against adversarial inputs. **Legal Engineering:** We call for a formalization of "Machine Readable IHL"—a digital Geneva Convention that can be parsed by autonomous agents. **Policy Standards:** We urge NATO and allied defense bodies to adopt the "Sentinel Standard" as a baseline for the certification of defensive autonomous systems.

In the end, the Sentinel is not a soldier. It has no honor, no courage, and no instinct for survival. It is a shield. And like any shield, its only purpose is to break so that the human behind it does not have to.

References

Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. *AI & Society*, 22(4):477–493.

Asimov, I. (1950). *I, Robot*. Gnome Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Brunstetter, D. and Braun, M. (2013). From jus ad bellum to jus ad vim: Recalibrating our understanding of the moral use of force. *Ethics & International Affairs*, 27(1):87–106.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5:40–60.

International Committee of the Red Cross (1987). Commentary on the additional protocols of 8 june 1977 to the geneva conventions of 12 august 1949.

International Court of Justice (1986). Military and paramilitary activities in and against nicaragua (nicaragua v. united states of america), merits, judgment. I.C.J. Reports 1986, p. 14.

International Court of Justice (1996). Legality of the threat or use of nuclear weapons, advisory opinion. I.C.J. Reports 1996, p. 226.

NATO (2021). Summary of the nato artificial intelligence strategy. Brussels.

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.

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

U.S. Department of Defense (2023). Dod directive 3000.09: Autonomy in weapon systems. Washington, DC.

Walzer, M. (1977). *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. Basic Books.