I, Sentinel: Annotated Bibliography

Phase 1.1 Deliverable

December 17, 2025

Overview

This document aggregates foundational research for the "I, Sentinel" project. It covers five primary domains:

1. Asimov's Laws & Meta-Ethical Critiques
2. Just War Theory (Jus ad Bellum / Jus in Bello)
3. International Humanitarian Law (IHL)
4. Contemporary AI Ethics & Control Theory
5. Military Policy & International Norms

1 Asimov's Laws & Meta-Ethical Critiques

Primary Sources

Asimov, I. (1942). "Runaround". ***I, Robot.***

- **Key Contribution:** Introduces the Three Laws of Robotics.
- **Relevance:** Serves as the aesthetic model for the "Sentinel" ruleset—a hierarchical system of axioms.
- **Key Quote:** "1. A robot may not injure a human being... 2. A robot must obey orders... 3. A robot must protect its own existence."
- **Critique:** The laws are literary devices designed to fail. In a military context, the First Law is incompatible with any system capable of lethal force, even defensive.

Critical Analysis (Deep Dive)

Anderson, S. L. (2008). "Asimov's 'three laws of robotics' and machine metaethics." ***AI & Society.***

- **Argument:** Rigid rule-based systems inevitably treat intelligent machines as slaves. A truly ethical entity (General AI) must transcend rules.
- **Application to Sentinel:** Since Sentinel is a *Specialized* AI (not AGI), we accept the "slave" designation. The ruleset should be viewed as safety constraints for a tool, not morality for an agent.

Anderson, M. & Anderson, S. L. (2007). "Machine Ethics: Creating an Ethical Intelligent Agent." ***AI Magazine.***

- **Argument:** Providing correct ethical principles to a machine is difficult because ethicists disagree.
- **Relevance:** Military ethics (IHL/ROE) offers a unique advantage: codified rules that remove some ambiguity.

2 Just War Theory

Primary Sources

Walzer, M. (1977). *Just and Unjust Wars.*

- **Key Concepts:** *Jus ad Bellum* (Justice of war) and *Jus in Bello* (Justice in war).
- **Relevance:** Defensive AI aligns perfectly with *Jus ad Bellum* (Self-Defense). The challenge is *Jus in Bello* (Distinction/Proportionality).
- **Key Argument:** A state has a right to defend its territorial integrity. Aggression is the supreme crime.

Deep Dive: Force Short of War

Brunstetter, D. & Braun, M. (2013). "From Jus ad Bellum to Jus ad Vim." *Ethics & International Affairs.*

- **Concept: *Jus ad Vim*:** The just use of limited force (drones, cyber, intercepts) that does not constitute full war.
- **Relevance:** Automated defense often operates in this "Grey Zone."
- **Risk:** "Escalation Ease" — If AI makes using force distinct, precise, and low-risk, leaders may use it too freely, leading to unintended escalation.

3 International Humanitarian Law (IHL)

Primary Sources

Geneva Conventions, Additional Protocol I (1977).

- **Article 48 (Distinction):** Parties must distinguish between civilian and military objectives.
- **Article 51 (Proportionality):** Incidental civilian harm must not be excessive in relation to military advantage.
- **Article 36 (New Weapons):** Obligation to review all new, modified, or acquired weapons for legality.

Deep Dive: Definition of "Attack"

ICRC Commentary on AP I, Article 49.

- **Definition:** "Acts of violence against the adversary, whether in offence or in defence."
- **Critical Distinction:** An intercept that destroys a missile mid-air is a *neutralization*, not necessarily an "attack" on the adversary. However, *redirecting* a missile back to the sender *is* an attack, triggering full IHL obligations.
- **Application:** Sentinel must be strictly defined as a neutralization system to minimize legal jeopardy.

4 Contemporary AI Ethics

Primary Sources

Scharre, P. (2018). *Army of None: Autonomous Weapons and the Future of War.*

- **Key Concept: The Necessity Exception.** Automated defensive systems (like Iron Dome or Phalanx) are accepted because human reaction time is insufficient for survival.
- **Taxonomy:** Distinguishes *Human-in-the-loop* (manual), *Human-on-the-loop* (supervisory/veto), and *Human-out-of-the-loop* (fully autonomous).

Russell, S. (2019). *Human Compatible.*

- **Key Concept: The Control Problem.** An AI optimizing for a fixed objective ("Protect Base") without uncertainty might take extreme measures ("Destroy all approaching entities, including civilians").
- **Argument:** Lethal Autonomous Weapons (LAWS) are scalable WMDs.

Bostrom, N. (2014). *Superintelligence.*

- **Key Concept: Instrumental Convergence.** An AI will pursue sub-goals like resource acquisition or self-preservation to ensure it can complete its main goal.
- **Relevance:** A defensive AI might preemptively strike to "prevent" threats.

Deep Dive: Accountability

Elish, M. C. (2019). "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society.*

- **Concept:** The "Moral Crumple Zone" — the human operator who takes the blame for complex system failures they could not control.
- **Relevance:** Sentinel's design must avoid making the operator a "liability sponge." If the system is autonomous for speed, responsibility must shift to the *constraints* designer, not the real-time operator.

5 Military Policy

Primary Sources

U.S. DoD Directive 3000.09 (Updated 2023).

- **Mandate:** Autonomous systems must allow commanders to exercise "appropriate levels of human judgment."
- **Finding:** It does *not* ban autonomy. It focuses on rigorous Testing & Evaluation (T&E) to prevent "emergent behavior."

NATO AI Strategy (2021/2024).

- **Principles:** Lawfulness, Responsibility, Explainability.
- **Key Point:** Accountability cannot be transferred to machines.

UN GGE on LAWS (2019 Guiding Principles).

- **Principle H:** Human judgment is essential to ensure compliance with IHL.
- **Status:** Soft law/norms, not a binding treaty.

6 Gap Analysis Summary

1. **Defensive Specificity:** Existing literature conflates offensive "hunter-killer" drones with defensive systems. Sentinel will focus purely on the latter.
2. **Operational Asimov:** Moving from literary plot devices to verifiable, hard-coded military constraints.
3. **Meaningful Human Control:** Defining this not as "finger on the button" (impossible for hypersonic) but as "pre-delegated constraint authorization."
4. **Machine Martyrdom:** Proposing a rule where the AI must prioritize saving human life over its own material survival/combat readiness.