Robots In War

Table of contents

- Robots In War
  - Table of contents
- Autonomous Driving Ethics: from Trolley Problem to Ethics of Risk
  - **1. Critique of the Trolley Problem and Transition to Risk Ethics**
  - **2. Limitations of Traditional Ethical Theories**
  - **3. Ethics of Risk: A Hybrid Framework**
  - **4. Application to the Trolley Problem**
  - **5. Social and Technical Implications**
  - **6. Unresolved Challenges and Future Directions**
  - **Conclusion**
- Just War and Robots' Killings
  - **1. Framing the Debate: Sparrow's Responsibility Trilemma**
  - **2. Rejecting the Trilemma: Engineering and Tolerance Levels**
  - **3. Ethics of Risk and the Precaution Thesis**
  - **4. Addressing Objections: Respect and Normalization**
  - **5. Practical Implications: Regulation Over Banning**
  - **6. Unresolved Tensions and Limitations**
  - **7. Synthesis with Just War Theory**
  - **8. Critical Examples and Analogies**
  - **9. Conclusion**
- Killer Robots by Robert Sparrow
  - **1. Core Argument: The Responsibility Trilemma**
  - **2. Ethical Foundations: *Jus in Bello* and Respect for Persons**
  - **3. The Child Soldier Analogy**
  - **4. Technological and Military Pressures**
  - **5. Unresolved Tensions and Counterarguments**
  - **6. Conclusion: Ethical and Policy Implications**

Autonomous Driving Ethics: from Trolley Problem to Ethics of Risk

1. Critique of the Trolley Problem and Transition to Risk Ethics

The paper begins by dismantling the trolley problem's relevance to autonomous vehicles (AVs), arguing that its **deterministic, binary framework** fails to capture real-world complexities. The authors highlight three critical shortcomings:

- **Certainty of Outcomes**: Trolley scenarios assume fatalities are inevitable, whereas real driving involves probabilistic risks. For example, collisions depend on sensor accuracy, prediction errors, and occluded objects (Fig. 4), making absolute certainty unrealistic.

- **Binary Choices vs. Continuous Solutions**: Unlike the trolley's two-track dilemma, AVs navigate a "continuous solution space for trajectories" (p. 3), requiring nuanced risk assessments across countless potential paths.
- **Lack of Context**: Trolley problems omit critical prior information (e.g., who caused the risk), which is ethically essential. Nyholm & Smids (2016) and Kauppinen (2020) emphasize that moral responsibility hinges on contextual factors like fault or intent, which the trolley framework ignores.

The authors pivot to **ethics of risk**, which evaluates actions under uncertainty. This shift aligns with Bonnefon et al. (2019), who reframe AV decisions as probabilistic risk distributions rather than life-and-death trade-offs. For instance, an AV adjusting its lateral position (Fig. 2) redistributes risk between cyclists and passengers based on collision probabilities and harm severity.

---

2. Limitations of Traditional Ethical Theories

The paper systematically critiques three ethical frameworks for AVs:

A. Deontology
Rule-based systems like Kant's Categorical Imperative or Asimov's Three Laws are deemed **too rigid** for dynamic environments. While Gerdes & Thornton (2015) propose hierarchical rules (e.g., prioritizing human safety over obedience), the authors argue such approaches:

- Fail to account for context (e.g., swerving to avoid a pedestrian might endanger others).
- Create conflicts between rules, leading to "dangerous behaviors" (p. 6) if rigidly enforced.
- Lack universality, as corner cases require endless rule additions.

B. Utilitarianism
Though utilitarian cost functions (minimizing total harm) are technically feasible and socially preferred (Bonnefon et al., 2016), they face **moral and legal challenges**:

- Sacrificing individuals for collective benefit violates human dignity, as per the German Ethics Commission (2017).
- Transparency issues arise when AVs prioritize "invisible" statistical lives over identifiable passengers (Hubner & White, 2018).
- Ignores fairness, as noted in Keeling's (2018) critique of Leben's (2017) Rawlsian algorithm, which disproportionately favors worst-off individuals.

C. Virtue Ethics
Machine learning (ML)-based approaches, like imitation learning (Bansal et al., 2018), aim to encode "virtues" such as prudence. However, the authors identify **critical flaws**:

- ML models lack explainability, making accountability impossible (Berberich & Diebold, 2018).
- Training data reflect common behaviors, not ethical ideals (Etzioni & Etzioni, 2017). For example, AVs might mimic aggressive human driving rather than virtuous caution.

---

3. Ethics of Risk: A Hybrid Framework

The authors propose a **risk-based model** integrating three decision principles:

1. **Bayesian Principle**: Minimize total risk (J_B = \sum R_i).
2. **Equality Principle**: Reduce risk disparities (J_E = \sum |R_i - R_j|).
3. **Maximin Principle**: Mitigate worst-case harm (J_M = \max(H_i)).

These are combined into a weighted cost function:
[J_{\text{total}} = (w_B J_B + w_E J_E + w_M J_M) \gamma^t
]
where (\gamma^t) discounts future risks.

Key Innovations:

- **Probabilistic Harm Quantification**: Risk (R = p(u)H(u)) incorporates collision probability ((p)) and harm severity ((H)), estimated via kinetic energy models (Sobhani et al., 2011). This avoids subjective valuations (e.g., monetizing lives) and focuses on physical metrics like speed and mass.
- **Fairness Through Hybridization**: By balancing total risk reduction (Bayesian), equity (Equality), and worst-case avoidance (Maximin), the framework addresses Keeling's (2018) critiques. For example:
  - In Fig. 5, the Equality Principle alone would favor two certain deaths over one minor risk, but combining it with Bayesian weights ((w_B > 0)) corrects this.
  - In Fig. 6, the Maximin Principle ignores cumulative harm to many, but adding Bayesian terms prioritizes collective safety.

Limitations:

- **Subjective Weightings**: The choice of (w_B, w_E, w_M) is unresolved. Manufacturers might prioritize passenger safety (high (w_M)), contradicting the German Ethics Commission's rejection of "sacrifice" logic.
- **Responsibility Metrics**: While the framework discounts risks temporally ((\gamma^t)), it doesn't penalize negligent road users (e.g., jaywalkers). Kauppinen (2020) argues culpability must influence risk distribution, but the paper defers this to future work.

---

4. Application to the Trolley Problem

The authors demonstrate their framework's flexibility by applying it to the trolley dilemma:

- **Standard Scenario**: Killing one vs. five. With (w_B > 0), Bayesian dominance ((J_B = 1) vs. (5)) yields the utilitarian outcome.
- **Modified Scenario**: Five suffer minor harm ((H = 0.2)) vs. one killed. Here, Maximin ((J_M = 0.2)) and Equality ((J_E = 1)) override Bayesian ((J_B = 1)), favoring the five.

This illustrates the framework's adaptability but also its dependence on weightings. The authors concede that without predefined weights, outcomes can be ambiguous, stating:

> "the question of the weighting factors [...] cannot be answered separately" (p. 17).

---

5. Social and Technical Implications

- **Mandatory vs. Personal Ethics**: The framework accommodates both approaches. Weights could be standardized (mandatory) or user-adjusted (personal "ethical knob" à la Contissa et al., 2017).

However, the authors warn that personal settings might increase insurance costs or societal risk.
- **Transparency and Regulation**: Explicit mathematical models enable post-accident audits, addressing the "black box" critique of ML-driven systems. Regulators could mandate transparency in (w) values to ensure compliance with ethical guidelines.
- **Cultural Variability**: The paper acknowledges but does not resolve how cultural differences in risk tolerance (e.g., individualistic vs. collectivist societies) might influence weightings.

---

6. Unresolved Challenges and Future Directions

- **Quantifying Responsibility**: The framework doesn't penalize at-fault road users. Future work might assign risk discounts based on culpability (e.g., red-light violators).
- **Dynamic Weightings**: Context-dependent (w) adjustments (e.g., higher (w_M) in school zones) could enhance fairness.
- **Legal Integration**: The German Ethics Commission's guidelines (2017) reject utilitarian sacrifices but permit probabilistic risk minimization. Aligning the framework with such policies requires legal-philosophical reconciliation.

Conclusion

Geisslinger et al. (2021) provide a groundbreaking shift from abstract trolley dilemmas to actionable risk ethics for AVs. Their hybrid framework balances competing ethical principles and operationalizes them into tractable mathematics. However, the reliance on subjective weightings and unresolved responsibility metrics leaves critical gaps. Future research must address these to ensure AVs navigate not just roads, but moral landscapes, with rigor and fairness. As the authors conclude:

> "The question of what constitutes a fair distribution of risk [...] should be at the center of future research" (p. 20).

Just War and Robots' Killings

---

1. Framing the Debate: Sparrow's Responsibility Trilemma

The paper opens by addressing Rob Sparrow's **responsibility trilemma**, a central ethical challenge to lethal autonomous weapons systems (LAWS). Sparrow argues that if a robot commits a war crime (e.g., targeting surrendering soldiers), there is no morally responsible agent:

- **Designers** are absolved because autonomous systems are designed to act unpredictably.
- **Commanders** cannot foresee autonomous actions, akin to artillery operators who are not blamed for misuse by soldiers.
- **Robots themselves** lack moral agency, as they cannot suffer or comprehend responsibility (pp. 304–305).

This creates a **responsibility gap**, rendering LAWS impermissible under Just War theory's requirement for accountability. Sparrow analogizes LAWS to child soldiers in "grey areas" of autonomy, where no party bears responsibility for atrocities (p. 305).

2. Rejecting the Trilemma: Engineering and Tolerance Levels

Simpson and Müller counter by introducing **tolerance levels**, a concept borrowed from engineering ethics. Tolerance levels define the reliability a system must achieve, integrating technical and moral considerations:

- **Example**: A bridge's collapse due to a 300-year flood event falls outside its tolerance level; no one is blameworthy. Conversely, failure due to poor design or misuse (e.g., overloading) assigns responsibility to engineers or users (pp. 307–309).
- **Application to LAWS**: If a robot operates within its tolerance level (e.g., targeting combatants with 99% accuracy), designers, commanders, or regulators are responsible for malfunctions. If it operates outside (e.g., unforeseeable sensor failure in extreme conditions), no blame accrues (pp. 310–311).

This framework rejects Sparrow's trilemma by distributing responsibility across a **chain of human actors**, similar to liability in civilian engineering projects.

3. Ethics of Risk and the Precaution Thesis

The authors pivot to the **ethics of risk imposition**, addressing whether LAWS can fairly redistribute risk. Key arguments include:

- **Risk Baselines**: The moral permissibility of LAWS hinges on whether they reduce non-combatant risk compared to human soldiers ((r_2 < r_1)). For example, if LAWS reduce collateral damage by 50%, their deployment is justified despite unequal risk distribution (pp. 314–315).
- **Policy G Analogy**: Reducing risk unequally (e.g., eliminating risk for soldiers while slightly reducing it for non-combatants) is acceptable if no egalitarian alternative exists. This aligns with Rawlsian prioritarianism, prioritizing absolute risk reduction over perfect equality (p. 315).

The **Precaution Thesis**—requiring precautions to minimize harm to each individual—is satisfied if LAWS meet two conditions:

1. (r_2 < r_1) (overall risk reduction).
2. (r_2) is minimized to the "technologically feasible" limit (p. 316).

4. Addressing Objections: Respect and Normalization

The paper confronts two objections:

- **Respect for Victims**: Sparrow and Nagel argue that LAWS violate interpersonal respect by depersonalizing killing. Simpson and Müller respond with **technological normalization**, comparing LAWS to artillery or drones:

  > "Familiarity with the weapon has rendered it no less disrespectful as a tool for killing than face-to-face combat" (p. 320).
  > Historical precedents (e.g., artillery) show that societal acceptance evolves, dissolving initial moral revulsion.

- **Psychological Aversion**: The "atavistic horror" of being killed by robots is likened to tax aversion—irrational but transient. Over time, LAWS would become routine, much like self-checkout systems (p. 318).

5. Practical Implications: Regulation Over Banning

The authors conclude with a pragmatic stance: **regulate, don't ban**. Key regulatory imperatives include:

- Setting **strict tolerance levels** for LAWS to ensure compliance with Just War principles (e.g., discrimination, proportionality).
- Licensing and testing regimes to enforce accountability (p. 320).
- International governance to prevent misuse by "badly ordered societies" (p. 320).

6. Unresolved Tensions and Limitations

- **Threshold Objection**: LAWS might lower the threshold for war by making conflict less politically costly (e.g., reduced soldier casualties). This risks increasing unjust wars (p. 320).
- **Democratic Accountability**: Centralized control of LAWS could bypass public consent, undermining democratic oversight (p. 320).
- **Liability in Asymmetric Conflicts**: The paper assumes LAWS will be used by "well-ordered societies," but real-world deployment (e.g., by non-state actors) complicates accountability.

7. Synthesis with Just War Theory

The authors align their argument with **revisionist Just War theory**, rejecting the "moral equality of combatants" and emphasizing liability based on threat contribution (pp. 310–311). They sidestep debates over combatant liability by focusing on risk redistribution, asserting that LAWS can satisfy *jus in bello* principles if calibrated to minimize non-combatant harm.

8. Critical Examples and Analogies

- **Mefloquine Case**: Pharmaceutical companies are not blamed for rare side effects if the drug meets regulatory tolerance levels (p. 309). Similarly, LAWS operators are blameless for statistically inevitable malfunctions.
- **Dam Construction**: A dam reducing flood risk, despite introducing a new collapse risk, is justified if $r_2 < r_1$ (p. 314).

9. Conclusion

Simpson and Müller provide a robust defense of LAWS within Just War theory, reframing responsibility through engineering ethics and risk analysis. However, their reliance on **technological optimism** (e.g., assuming $r_2 < r_1$) and societal normalization leaves room for critique. As they concede:

> "The deep and difficult question [is] whether it is likely that r_2 will be greater or lesser than r_1" (p. 316).

The paper's strength lies in its integration of moral philosophy with practical engineering standards, offering a blueprint for ethical LAWS deployment—**if** regulators enforce stringent tolerance levels and prioritize non-combatant safety.

Killer Robots by Robert Sparrow

1. Core Argument: The Responsibility Trilemma

Sparrow's central thesis revolves around the **responsibility gap** inherent in deploying lethal autonomous weapon systems (AWS). He argues that AWS's autonomy precludes meaningful accountability for war crimes, violating the ethical foundations of *jus in bello*. The trilemma posits three potential loci of responsibility—programmers, commanders, and machines—and systematically dismantles each:

1. **Programmers**:
    - **Argument**: Programmers cannot foresee all actions of a learning, adaptive AWS. Autonomy breaks the causal chain between programming and outcomes.
    - **Example**: A robot that evolves beyond its initial code to target surrendering soldiers (p. 66) cannot have its designers blamed, as its decisions reflect "internal states" (beliefs, desires) shaped post-deployment.
    - **Limitation**: Assumes programmers cannot implement fail-safes or ethical constraints. Contemporary debates on "value alignment" in AI challenge this (e.g., embedding Asimov's laws).
2. **Commanding Officers**:
    - **Argument**: Commanders are likened to artillery operators, responsible for deployment but not specific targeting. However, AWS's unpredictability makes this analogy flawed.
    - **Quotation**: "The more autonomous the systems are, the larger this risk looms. At some point, it will no longer be fair to hold the Commanding Officer responsible" (p. 70).
    - **Counterpoint**: Military doctrine already holds commanders accountable for collateral damage from "dumb" weapons (e.g., artillery). Why not extend this to AWS? Sparrow dismisses this by emphasizing AWS's unique decision-making capacity (p. 70).
3. **Machines Themselves**:
    - **Argument**: Machines lack moral agency. Punishing them is nonsensical, as they cannot "suffer" or comprehend guilt (p. 72).
    - **Analogy**: Assigning blame to a machine is as absurd as prosecuting a malfunctioning toaster (p. 71).
    - **Philosophical Challenge**: Sparrow presupposes a Kantian view of moral responsibility requiring consciousness. Proponents of "artificial moral agents" (Floridi & Sanders, 2001) argue responsibility could hinge on functional behavior, not sentience.

2. Ethical Foundations: *Jus in Bello* and Respect for Persons

Sparrow grounds his argument in Just War Theory's requirement for accountability:

- **Key Principle**: "The least we owe our enemies is allowing that their lives are of sufficient worth that someone should accept responsibility for their deaths" (p. 67).
- **Consequences**: Unaccountable AWS reduce warfare to "extermination," akin to using landmines or WMDs (p. 67).

Strengths:

- Highlights the dehumanizing effect of detached, automated killing.
- Aligns with Nagel's (1972) emphasis on interpersonal respect in war: Combatants deserve to know *who* decided their fate and *why*.

Weaknesses:

- **Technological Determinism**: Assumes AWS will inherently lack oversight. Yet, hybrid systems (e.g., human-AI collaboration) could retain accountability (e.g., requiring final human approval for strikes).
- **Historical Precedent**: Drones already operate with human oversight. Sparrow acknowledges this but argues autonomy trends will eliminate "the human in the loop" (p. 68–69).

3. The Child Soldier Analogy

Sparrow's most provocative comparison links AWS to child soldiers:

- **Shared Trait**: Both occupy a "grey area" of partial autonomy. Children lack full moral agency; AWS lack moral personhood.
- **Ethical Dilemma**: "No one is in control. If civilians are killed, they are killed senselessly without anyone being responsible" (p. 73).
- **Implication**: Using AWS mirrors the moral bankruptcy of deploying child armies (e.g., Liberia, Angola).

Critique:

- **False Equivalence**: Child soldiers are victims coerced into violence, whereas AWS are tools designed by humans. The analogy conflates *moral patients* (children) with *amoral instruments* (robots).
- **Policy Response**: International law bans child soldiers via the Optional Protocol to the CRC. A similar ban on AWS (e.g., proposed UN Treaty) could resolve Sparrow's trilemma.

4. Technological and Military Pressures

Sparrow identifies systemic drivers pushing toward AWS deployment:

1. **Tempo of Battle**: AI's speed outperforms human decision-making in air combat (p. 68).
2. **Cost-Efficiency**: AWS reduce soldier casualties and financial burdens (p. 64).
3. **Survivability**: Removing human operators mitigates communication vulnerabilities (p. 69).

Contradiction: While AWS promise precision (reducing collateral damage), their autonomy risks *increased* unpredictability. Sparrow cites the LOCAAS missile system, which autonomously selects warhead configurations (p. 64), as a harbinger of ethical chaos.

5. Unresolved Tensions and Counterarguments

- **Regulation vs. Ban**: Sparrow dismisses regulation ("Human oversight will eventually be eliminated," p. 68) but ignores frameworks like the EU's AI Act, which mandates human control over high-risk AI.
- **Moral Progress**: If AWS reduce civilian casualties compared to human soldiers, their use might be *more* ethical despite accountability gaps. Sparrow's focus on responsibility overlooks consequentialist gains.
- **Moral Agency Evolution**: Advances in AI consciousness (e.g., artificial general intelligence) could render AWS "moral persons." Sparrow rejects this (p. 72) but does not engage with futurists like Kurzweil (1999).

6. Conclusion: Ethical and Policy Implications

Sparrow’s analysis remains a cornerstone in AWS ethics, compellingly arguing that autonomy erodes accountability. However, his conclusions face challenges:

- **Techno-Optimism**: Engineers like Arkin (2009) propose "ethical governors" to constrain AWS behavior, potentially resolving responsibility gaps.
- **Legal Precedent**: The Ottawa Treaty banned landmines; a similar ban on AWS could preempt Sparrow’s dystopia.
- **Philosophical Evolution**: Debates on AI moral agency (e.g., machine consciousness) may redefine responsibility paradigms.

Ultimately, Sparrow’s warning—that AWS risk rendering war "unfair either to potential casualties [...] or to the officer who will be held responsible" (p. 74)—underscores the urgency of ethical and legal frameworks to govern autonomous weapons before they become battlefield mainstays.