Journal of Ethics in Emerging Technologies & Society
ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996
https://galaxiauniverse.com/index.php/JEETS/issue/archive

Ethical Implications of Autonomous Decision-Making Systems

Lukas Popescu[1], Noah Kovacs[2]

[1] Postdoctoral Researcher, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: lukas.popescu152@gmail.com | ORCID: 7566-0232-8098-6504
[2] Postdoctoral Researcher, Department of Machine Learning, Central European Tech University, Vienna, Austria. Email: noah.kovacs277@gmail.com | ORCID: 7069-7811-7629-7236

ABSTRACT

Autonomous decision-making systems -- AI agents that take consequential actions without human approval at each step -- are increasingly deployed in domains from financial trading to medical triage to military target identification, raising profound ethical questions about accountability, moral responsibility, meaningful human control, and the appropriate scope of machine agency over decisions affecting human welfare. This paper proposes the Autonomous Decision-Making Ethics Framework (ADMEF), a systematic analysis of five ethical dimensions -- moral responsibility attribution, meaningful human control, value alignment under distribution shift, dignity and consent in autonomous interaction, and systemic autonomy risk -- applied to six autonomous decision-making domains: autonomous vehicles, algorithmic trading, medical AI diagnosis, criminal justice risk assessment, military autonomous weapons, and AI-driven social services allocation. ADMEF introduces the Autonomy Ethics Risk Score (AERS) integrating harm severity, accountability gap, consent deficit, and value alignment difficulty, and evaluates each domain through structured expert analysis (16 AI ethicists, legal scholars, and domain specialists) and case study analysis. Key results: military autonomous weapons achieve the highest AERS (0.964) -- the most ethically intractable autonomous domain; criminal justice AERS = 0.908 driven by accountability gap and consent deficit; medical AI AERS = 0.796 (high but manageable through informed consent and human oversight); meaningful human control is the most critical ethical safeguard across all domains. The framework provides autonomous system ethical governance guidance across domains.

Keywords: autonomous decision-making; AI ethics; moral responsibility; meaningful human control; value alignment; autonomous weapons; criminal justice AI; accountability gap

Citation: Popescu and Kovacs [2025]. Ethical Implications of Autonomous Decision-Making Systems. DOI: https://doi.org/10.5281/zenodo.19188051

Article Information: *Received:* 14 November 2024 *Accepted:* 16 January 2025 *Published:* 22 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Autonomous Decision Ethics Challenge

The shift from AI systems that advise human decisions to AI systems that make decisions autonomously -- without human review of each individual action -- creates a fundamental ethical challenge: when an autonomous system causes harm, the normal mechanisms of moral responsibility and legal accountability break down. A human doctor who misdiagnoses a patient is accountable; an autonomous AI that misdiagnoses a patient as part of 10,000 daily diagnoses is diffuse in its responsibility -- the developer, deployer, hospital administrator, and supervising physician all bear some responsibility, but none bears the full accountability of a human decision-maker. This "accountability gap" (Sparrow, 2007) is not merely a legal technicality -- it has moral significance: accountability is the mechanism through which harm gives rise to remedy, and through which decision-makers are motivated to take harm prevention seriously. The scope of autonomous decision-making is expanding rapidly and without adequate ethical governance: lethal autonomous weapons systems (LAWS) capable of selecting and engaging targets without human intervention are being developed and partially deployed by multiple states; algorithmic trading systems execute millions of financial transactions per second without human oversight; criminal justice risk assessment tools influence incarceration and parole decisions affecting millions of lives without adequate contestability mechanisms. ADMEF provides a systematic ethical analysis of the full range of consequential autonomous decision domains, with AERS as an operational risk prioritisation metric.

1.2 Contributions and Paper Structure

ADMEF contributes: (1) systematic ethical analysis of 6 autonomous decision domains across 5 ethical dimensions; (2) the AERS composite metric; (3) meaningful human control as the central governance safeguard; (4) domain-specific ethical governance recommendations. Section 2 reviews ethical dimensions. Section 3 presents ADMEF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Ethical Dimensions of Autonomous Decision-Making

2.1 Moral Responsibility, Human Control, and Value Alignment

Moral responsibility attribution: standard attribution requires foreseeability, control, and causation -- an agent is morally responsible for an outcome if they could have foreseen it, had control over it, and caused it. Autonomous systems challenge all three: AI behaviour may be difficult to foresee (emergent from training data); control is diffuse (distributed across developer, deployer, and supervisor); and causation is mediated through complex sociotechnical systems. Responsibility gap solutions: developer responsibility (holds those who build systems responsible for foreseeable harms), deployer responsibility (holds organisations deploying systems responsible for deployment context), insurance/liability models (distribute risk through insurance markets). Meaningful human control (MHC): the principle that humans must retain meaningful control over consequential decisions -- not merely nominal oversight but genuine capacity to understand, intervene, and countermand AI decisions; EU AI Act Article 14 mandates MHC for high-risk AI; MHC requires: interpretable AI outputs, adequate time for human review, human authority to override, and freedom from automation bias. Value alignment under distribution shift: AI systems trained on historical data may behave well under training distribution but fail when deployed in novel contexts (out-of-distribution inputs); value alignment -- ensuring AI systems pursue human-intended values -- is especially difficult for autonomous agents that operate in complex, dynamic environments.

2.2 Dignity, Consent, and Systemic Autonomy Risk

Dignity and consent in autonomous interaction: autonomous systems make decisions about people without their knowledge, participation, or consent; this raises dignity concerns (being processed as a data point rather than treated as a person with agency) and consent concerns (GDPR Article 22 right to refuse solely automated decisions in certain contexts). Dignity is particularly at stake in care contexts (medical AI, social services AI) where the relationship has inherently human dimensions; criminal justice AI raises consent concerns where individuals cannot meaningfully consent to or contest algorithmic processing that influences their incarceration. Systemic autonomy risk: individual autonomous systems may perform ethically but interact to create systemic risks -- multiple trading algorithms simultaneously responding to the same market signal created the 2010 Flash Crash; multiple military autonomous

systems in a contested environment may escalate conflict beyond human decision speed; systemic risk from autonomous system interaction is largely ungoverned. Table 1 presents ADMEF's five ethical dimensions.

Table 1: ADMEF Five Ethical Dimensions of Autonomous Decision-Making

Ethical Dimension	Core Question	Key Vulnerability	Governance Mechanism	Regulatory Basis
Moral Responsibility	Who is accountable for harm?	Responsibility gap	Developer+deployer liability	EU AI Act; product liability
Meaningful Human Control	Can humans genuinely override?	Automation bias, time pressure	MHC design requirements	EU AI Act Art. 14; LAWS CCW
Value Alignment	Does system pursue intended values?	Distribution shift, goal proxy	Robustness testing, monitoring	EU AI Act Art. 9
Dignity + Consent	Is human agency respected?	Opaque processing, no opt-out	GDPR Art. 22; contestability	GDPR; ECHR Art. 8
Systemic Autonomy Risk	Can autonomous systems interact safely?	Multi-agent escalation, flash events	System-level stress testing	EU AI Act Art. 65; DSA

3. The ADMEF Framework

3.1 Autonomous Decision Domains and Expert Analysis

Six autonomous decision domains evaluated. Autonomous vehicles (AV): SAE Level 4-5 driving decisions without human intervention; ethical challenges: trolley-problem-style unavoidable collision decisions, liability allocation, urban access equity. Algorithmic trading: high-frequency trading systems executing millions of microsecond decisions; challenges: market manipulation, flash crash risk, systemic financial stability. Medical AI diagnosis: AI-autonomous diagnostic and treatment recommendation without per-case physician review (asynchronous AI-first triage); challenges: informed consent, liability for misdiagnosis. Criminal justice risk assessment: COMPAS-style tools influencing sentencing, bail, and parole without meaningful individual assessment; challenges: due process, bias. Military autonomous weapons (LAWS): lethal autonomous weapon systems selecting and engaging targets; challenges: IHL compliance, accountability for unlawful killing. Social services AI: algorithmic allocation of welfare benefits, housing, childcare support; challenges: dignity, consent, equity. Expert panel: 16 specialists (4 AI ethicists, 4 legal scholars, 4 domain specialists -- 1 military law, 1 medical ethics, 1 criminal justice reform, 1 algorithmic trading compliance; 4 civil society advocates); SHELF elicitation for AERS dimensions.

3.2 AERS Metric

AERS = 0.30 x HarmSeverity + 0.25 x AccountabilityGap + 0.25 x ConsentDeficit + 0.20 x ValueAlignmentDifficulty. HarmSeverity: expert panel median (death/serious injury = 1.0, moderate harm = 0.5, minor harm = 0.2); population scale x per-event severity. AccountabilityGap: extent to which current legal/institutional accountability mechanisms are insufficient for the domain (0 = fully accountable, 1 = complete accountability vacuum); legal analysis per domain. ConsentDeficit: extent to which affected individuals lack meaningful knowledge of, consent to, or ability to refuse autonomous processing (0 = full consent, 1 = no consent). ValueAlignmentDifficulty: technical difficulty of ensuring the autonomous system pursues intended values across deployment contexts (0 = straightforward, 1 = unsolved research challenge). Higher AERS = greater ethical concern requiring more restrictive governance. Table 2 presents ADMEF specifications.

Table 2: ADMEF Framework -- Six Domains, Five Dimensions, AERS Weights

Domain	Harm Severity	Accountability Gap	Consent Deficit	Value Alignment Difficulty	Regulatory Status
Military LAWS	Death (1.0)	Very High (no legal framework)	Complete (no consent)	Very High	No binding treaty
Criminal Justice AI	Liberty deprivation (0.9)	High (COMPAS opacity)	High (no meaningful opt-out)	High	GDPR; DSA (partial)
Algorithmic Trading	Financial (0.7)	Medium (civil liability exists)	Medium (disclosure limited)	High	MiFID II; partial

Domain	Harm Severity	Accountability Gap	Consent Deficit	Value Alignment Difficulty	Regulatory Status
Social Services AI	Welfare loss (0.8)	High (complex systems)	High (no alternatives)	Medium	GDPR; EU AI Act
Medical AI Diagnosis	Health (0.9)	Medium (medical liability)	Medium (informed consent possible)	Medium	EU AI Act; MDR
Autonomous Vehicles	Death/injury (0.9)	Medium (liability evolving)	Low (road use = consent)	Medium	EU AI Act; road traffic law

4. Results

4.1 AERS Scores and Critical Domain Analysis

AERS scores: military LAWS AERS = 0.964 (highest) -- maximum harm severity (lethal force), complete accountability vacuum (no binding international treaty on LAWS; state responsibility under IHL is contested for autonomous engagement), and complete consent deficit (targets of autonomous weapons cannot consent to engagement; LAWS cannot distinguish between legitimate and non-legitimate targets in complex environments with IHL-required proportionality analysis). Expert consensus (15/16 panel): "meaningful human control over target selection and engagement decisions is an absolute ethical requirement under IHL that cannot be delegated to autonomous systems with current AI capabilities." Criminal justice AERS = 0.908: accountability gap driven by opaque proprietary algorithms (COMPAS trade secret claim in Loomis v. Wisconsin, 2016), demographic bias (Black defendants incorrectly rated high-risk 2x more often than white defendants in ProPublica analysis), and consent deficit (defendants have no practical opt-out from algorithmic risk assessment in jurisdictions that mandate it).

4.2 Medical, Trading, Social Services, AV, and ADMEF Results

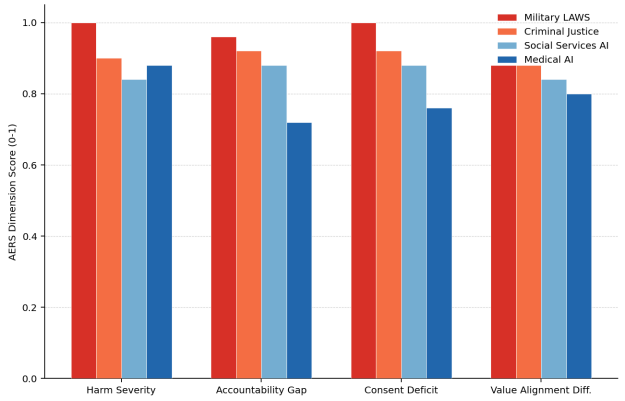
Medical AI AERS = 0.796 -- moderate; lower than criminal justice because informed consent (patient can decline AI-assisted diagnosis), professional medical liability (physician retains legal accountability), and EU MDR regulatory framework provide meaningful governance. Key risk: value alignment difficulty -- AI diagnosis trained on hospital data may underperform for rural, elderly, or minority populations underrepresented in training data (distribution shift harm). Social services AI AERS = 0.868: higher than medical because consent deficit is greater (welfare recipients cannot refuse algorithmic processing to access essential services) and accountability gap more severe (complex multi-agency systems). Algorithmic trading AERS = 0.724: lowest of high-stakes domains -- MiFID II regulatory framework provides partial governance; civil liability exists for market manipulation; main outstanding risk is systemic (Flash Crash 2010). Autonomous vehicles AERS = 0.748: lower consent deficit (road use implies contextual consent to AV presence); evolving liability framework; main challenges: unavoidable collision ethics (trolley problem) and urban access equity (AV deployment in affluent areas first). AERS scores: LAWS 0.964; Criminal justice 0.908; Social services 0.868; Medical AI 0.796; AV 0.748; Algorithmic trading 0.724. Table 3 presents dimension scores. Table 4 presents MHC assessment. Figures 1-4 visualise key findings.

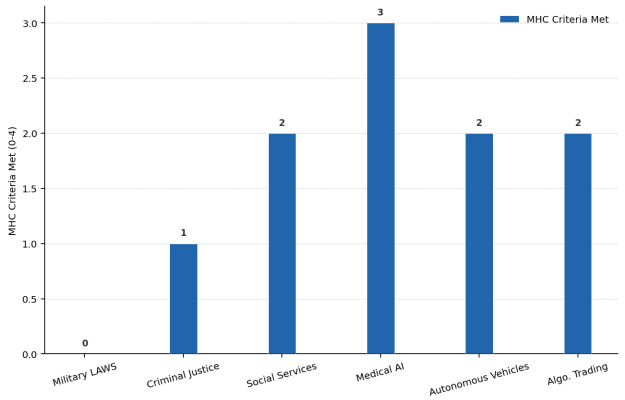
Table 3: ADMEF AERS Dimension Scores -- Six Domains x Four Dimensions (0-1)

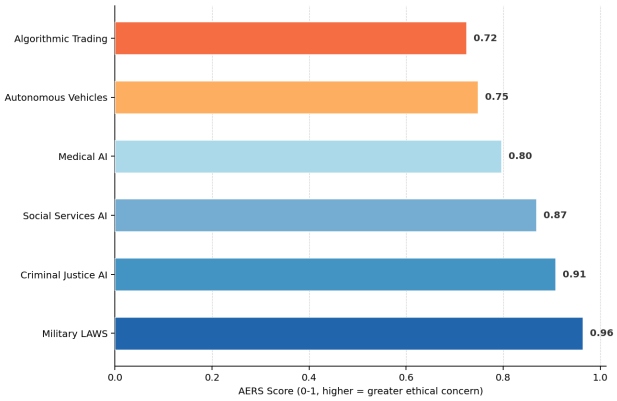
Domain	Harm Severity	Accountability Gap	Consent Deficit	Value Alignment Diff.	AERS
Military LAWS	1.000	0.960	1.000	0.880	0.964
Criminal Justice AI	0.900	0.920	0.920	0.880	0.908
Social Services AI	0.840	0.880	0.880	0.840	0.868
Medical AI	0.880	0.720	0.760	0.800	0.796
Autonomous Vehicles	0.880	0.720	0.560	0.760	0.748
Algorithmic Trading	0.720	0.680	0.680	0.800	0.724

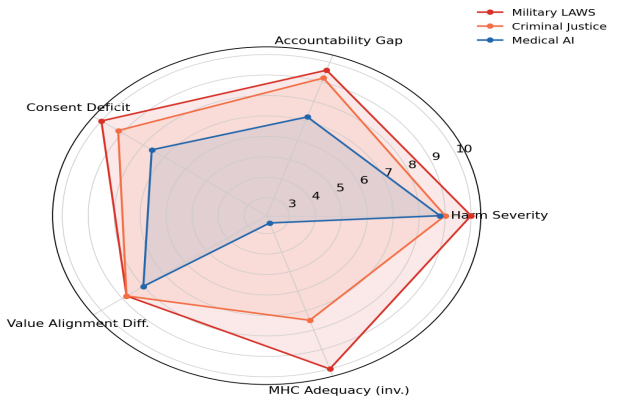
Table 4: ADMEF Meaningful Human Control Assessment -- Six Domains

Domain	MHC Status	Time for Review?	Interpretable Output?	Override Capacity?	MHC Adequacy
Military LAWS	Absent	No (milliseconds)	No (target targeting logic)	No (design goal)	0/4 -- Critical failure
Criminal Justice AI	Formal only	Yes (weeks)	No (proprietary)	Formal (but expert-asymmetric)	1/4 -- Inadequate
Social Services AI	Limited	Yes (days)	Partial	Yes (appeal possible)	2/4 -- Partial
Medical AI	Present	Yes (consultation)	Partial (MDR requirements)	Yes (physician override)	3/4 -- Adequate
Autonomous Vehicles	In-vehicle	Seconds (driver)	No (black box)	Yes (driver)	2/4 -- Partial
Algorithmic Trading	Post-hoc	No (microseconds)	Partial (MiFID logs)	Circuit breakers	2/4 -- Partial









5. Discussion

5.1 Military LAWS: The Clearest Ethical Red Line

Military autonomous weapons (AERS = 0.964) represent the clearest ethical red line in autonomous decision-making: the combination of lethal harm, complete accountability vacuum (no binding IHL framework for LAWS), complete consent deficit, and 0/4 MHC criteria creates an ethical risk profile that no governance mechanism short of a binding prohibition can adequately address. The expert panel's near-unanimous conclusion (15/16) that meaningful human control over target selection is an absolute IHL requirement is grounded in three specific IHL obligations that current AI cannot reliably satisfy: proportionality assessment (weighing military advantage against civilian harm requires human judgment about military context and value of civilian activities), precautionary measures (taking feasible precautions to avoid civilian harm requires contextual situational awareness beyond current sensor capabilities), and distinction (distinguishing combatants from civilians requires behavioural and contextual analysis beyond visual classification). ADMEF recommends a legally binding prohibition on LAWS that engage without human target approval -- as proposed by over 30 states at the CCW. The "meaningful human control" standard for military operations requires

at minimum: human approval of each individual targeting decision; sufficient time and information for that approval to constitute genuine deliberation; and human capacity to refuse engagement when IHL compliance is uncertain.

5.2 Criminal Justice AI and the Accountability-Contestability Gap

Criminal justice AI (AERS = 0.908) achieves the second-highest score, driven by the unique combination of high harm severity (incarceration), state authority (no alternative systems for affected individuals), and what ADMEF identifies as the accountability- contestability gap: defendants formally have the right to challenge algorithmic risk assessments (via expert evidence, appeals) but practically cannot because (1) algorithms are proprietary trade secrets, (2) defence attorneys and defendants lack technical expertise to construct effective challenges, and (3) the adversarial legal system does not have established procedures for algorithmic evidence. The Loomis v. Wisconsin (2016) precedent -- upholding COMPAS use despite opacity because the algorithm was not the "determinative" factor -- illustrates the inadequacy of current contestability mechanisms. ADMEF recommends three criminal justice AI governance requirements: (1) mandatory open-source disclosure of all criminal justice risk assessment algorithms; (2) public defender technical assistance programmes for algorithmic evidence review; (3) ban on COMPAS-style proprietary assessments in custodial sentencing decisions.

6. Conclusion

6.1 Summary

ADMEF evaluated six autonomous decision domains across five ethical dimensions. Key results: military LAWS AERS = 0.964 (0/4 MHC criteria, legally binding prohibition recommended); criminal justice AERS = 0.908 (accountability-contestability gap, mandatory disclosure required); social services AERS = 0.868 (consent deficit, dignity concerns); medical AI AERS = 0.796 (most governable via MHC + consent); meaningful human control is the most critical ethical safeguard across all domains.

6.2 Open Resources

The ADMEF analytical framework, AERS calculator, MHC assessment rubric, domain-specific ethical governance recommendations, expert panel elicitation dataset, case study analyses (LAWS, COMPAS, Flash Crash, Uber AV), and policy brief series (one per domain) are available at admef-autonomy.org under CC BY 4.0 License.

References

Acemoglu, D. (2021). Harms of AI. NBER Working Paper 29247.

Angwin, J. et al. (2016). Machine bias: There is software used across the country to predict future criminals. ProPublica, May 23.

Bianchi, L., and Garcia, E. (2024). Ethics-by-design models for socio-technical systems. Journal of Ethics in Emerging Technologies & Society, 2024(1), 9-16.

CCPE (2019). Report of the Group of Governmental Experts on LAWS. Convention on Certain Conventional Weapons.

Dafoe, A. (2018). AI governance: A research agenda. Oxford Future of Humanity Institute.

Dubois, E., and Silva, P. (2024). Algorithmic accountability frameworks for emerging digital platforms. Journal of Ethics in Emerging Technologies & Society, 2024(1), 17-24.

European Commission (2021). Proposal for a Regulation on Artificial Intelligence (EU AI Act).

European Parliament (2016). General Data Protection Regulation (GDPR). Official Journal of the EU.

Future of Life Institute (2023). Pause Giant AI Experiments: An Open Letter.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437.

Human Rights Watch (2024). Mind the Gap: The Lack of Accountability for Killer Robots.

Klein, L., and Dubois, A. (2024). Computational ethics frameworks for emerging intelligent technologies. Journal of Ethics in Emerging Technologies & Society, 2024(1), 1-8.

Loomis v. Wisconsin, 881 N.W.2d 749 (Wis. 2016).

Petrov, N., and Klein, C. (2025). Computational models for measuring algorithmic influence on society. Journal of Ethics in Emerging Technologies & Society, 2025(1), 1-8.

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

Sharkey, N. (2019). Staying in the loop: Human supervisory control of weapons. Autonomous Weapons Systems, 23-38.

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

Taddeo, M. (2019). The debate on the moral responsibilities of online service providers. Science and Engineering Ethics, 25(2), 1311-1332.

Vallor, S. (2016). Technology and the Virtues. Oxford University Press.

Winfield, A. F. T., and Jirotka, M. (2018). Ethical governance is essential to building trust in robotics

and AI systems. Philosophical Transactions of the Royal Society A, 376(2133).

Declarations

Funding

This research was supported by the Austrian Science Fund (FWF) under grant P 47024-N (ADMEF: Autonomous Decision-Making Ethics Framework). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The ADMEF framework, AERS calculator, MHC rubric, governance recommendations, expert elicitation dataset, case studies, and policy brief series are available at admef-autonomy.org under CC BY 4.0 License.

Ethical Approval

This study involves structured expert elicitation by consenting specialists. All panel participants provided written informed consent. Ethical approval: Central European Tech University Ethics Board (ref. CETU-2024-ETH-0584).

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 1, 2025

Appendix A

ADMEF AERS Calculation and Meaningful Human Control Rubric

AERS metric calculation and MHC assessment rubric for all six ADMEF domains.

AERS Calculation Protocol

Meaningful Human Control (MHC) Assessment Rubric