The American Journal of Interdisciplinary Innovations and Research
ISSN 2642-7478 | Open Access

Type Original Research
PAGE NO.111-118

Check for updates

Ethical Architectures for Autonomous Driving: Reconciling Trolley Problem Thought Experiments with Practicable Decision-Making Frameworks

Dr. Esteban R. Moreno
Global Institute for Technology and Ethics, University of Lisbon

OPEN ACCESS

SUBMITTED 01 August 2025
ACCEPTED 15 August 2025
PUBLISHED 31 August 2025
VOLUME Vol.07 Issue 08 2025

CITATION
Dr. Esteban R. Moreno. (2025). Ethical Architectures for Autonomous Driving: Reconciling Trolley Problem Thought Experiments with Practicable Decision-Making Frameworks. The American Journal of Interdisciplinary Innovations and Research, 7(8), 111–118. Retrieved from https://theamericanjournals.com/index.php/tajiir/article/view/7009

Abstract- This paper examines the persistent tension between abstract moral thought experiments—most notably the trolley problem—and the concrete engineering, legal, and social realities of autonomous vehicle decision-making. We synthesize philosophical analyses, empirical studies, technical descriptions of perception and control systems, and recent regulatory and dataset-auditing work to produce a coherent account of ethically bounded architectures for autonomous driving systems. The central argument is that a purely philosophical framing (e.g., sacrificial dilemmas) is insufficient for designing viable autonomous driving ethics; instead, ethically bounded AI must integrate formal ethical constraints, decision-theoretic planning, robust perception, human-centered accountability mechanisms, dataset auditing, and meaningful human control. We propose a layered framework that links (1) low-level safety and collision mitigation algorithms, (2) intermediate decision-theoretic planning incorporating probabilistic risk assessment and distributive ethical constraints, and (3) high-level governance mechanisms for transparency, auditing, and legal alignment. We analyze normative trade-offs between utilitarian, deontological, and contractualist approaches and show how these philosophical families map onto technical choices in architecture, datasets, and evaluation. The paper also explores responsibility attribution, the role of simulated

and real-world datasets in shaping moral behavior, and the procedural mechanisms needed to operationalize "meaningful human control." Finally, we identify research priorities and policy recommendations to move from rhetorical trolley-problem debates to implementable, justifiable systems that can be audited and regulated. The conclusions emphasize layered safeguards, multidisciplinary governance, dataset quality assurance, and a move away from binary sacrificial framing toward continuous harm-minimization under uncertainty. (max 400 words)

Keywords: Autonomous vehicles, ethics, trolley problem, decision-theoretic planning, dataset auditing, meaningful human control

Introduction

The advent of autonomous vehicles (AVs) has resuscitated classic moral thought experiments for a new technological era. The trolley problem—choosing between actions that sacrifice one life to save several—has become shorthand for the ethical quandaries engineers, ethicists, and policymakers confront when programming machines that may have to make life-and-death choices in fractions of a second (Zhao & Li, 2020; Wu, 2020). This focus on sacrificial dilemmas has generated intense public interest and academic debate, but it has also produced misconceptions about the nature of ethical decision-making in real-world AV systems. What appears as a neat, binary ethical problem in philosophy turns out to be a multi-dimensional design and governance challenge when confronted with noisy sensors, uncertain predictions, distributed liability, and societal expectations (Sven & Smids, 2016; Wendell & Colin, 2008).

Philosophical contributions—ranging from classical formulations of the doctrine of double effect (Foot, 1967) to contemporary analyses of killing versus letting die (Thomson, 1976)—provide indispensable conceptual tools for clarifying moral distinctions (Philippa, 1967; Judith, 1976). Yet, translating those distinctions into algorithms raises practical questions about feasibility, traceability, and societal legitimacy (Derek, 2017; Gray, 2012). Moreover, the technological substrate of AVs—deep perception stacks, decision-making modules, and control systems—is not neutral. Sensor limitations, prediction uncertainty, and the data used to train models all shape the set of feasible actions and therefore the moral choice space (Zhang et al., 2018; Keqiang, 2017). At the same time, concerns about transparency, accountability, and public trust necessitate governance instruments such as dataset audits, explainability efforts, and legal harmonization (European Parliament, 2022; Williams et al., 2022).

This paper addresses the gap between philosophical abstraction and engineering practice by offering a structured, multidisciplinary approach to ethically bounded decision-making in AVs. We document and analyze how moral theories map onto algorithmic choices; how perception, planning, and control constraints reshape ethical trade-offs; and what governance mechanisms can ensure accountability. In doing so, we argue for a layered ethical architecture in which low-level safety mechanisms work in concert with mid-level decision-theoretic planners and high-level governance processes. This architecture is grounded in the twin goals of continuous harm-minimization and societal legitimacy rather than in sensationalized sacrificial scenarios.

The literature gap motivating this study is threefold. First, while many papers critique trolley-problem framing for AVs, fewer works systematically connect that critique to concrete architectural alternatives that are implementable under realistic sensing and control constraints (Sven & Smids, 2016; Zhao & Li, 2020). Second, there is limited integration between technical literatures on planning and control and philosophical literatures on moral theories—a gap that obscures how ethical choices translate into algorithmic constraints (Basye et al., 1992; Dellermann et al., 2021). Third, recent policy and auditing work has not been fully synthesized with the ethical and technical literature to create operational recommendations for dataset and system governance (European Parliament, 2022; Rossi & Mattei, 2019). This article aims to bridge these gaps with a detailed theoretical elaboration and a prescriptive, layered architecture that is sensitive to legal, social, and technical realities.

Methodology

This paper uses an integrative, conceptual-methodological approach. Rather than reporting original empirical field data, the study synthesizes philosophical argumentation, engineering literature on perception and control, decision-theoretic planning research, and policy-oriented analyses. The methodology consists of four interlocking components: (1) conceptual mapping of moral theories to algorithmic primitives; (2) technical analysis of perception-planning-

control constraints and their ethical implications; (3) evaluation of governance instruments including dataset auditing and accountability mechanisms; and (4) construction of a layered ethical architecture and hypothetical scenario analyses to illustrate operational application.

Conceptual mapping begins with a systematic review of canonical moral frameworks—utilitarianism, deontology, virtue ethics, doctrine-of-double-effect reasoning—and identifies the decision-theoretic primitives each framework emphasizes (e.g., outcome maximization, rule-following, intention-focused constraints). We then map these primitives onto algorithmic constructs such as objective functions, hard constraints, reward shaping, and rule-based overrides. This mapping is informed by contemporary normative modeling work which demonstrates how philosophical stances can be represented in computational systems (Derek, 2017; Hennig & Hütter, 2020).

The technical analysis reviews core elements of AV stacks—sensor fusion, perception models, trajectory prediction, planning algorithms, and control—and identifies sources of uncertainty and failure modes (Zhang et al., 2018; Keqiang, 2017; Basye et al., 1992). We analyze how these technical constraints narrow the feasible action set and thus limit the range of ethically justifiable options. This analysis draws on literature from decision-theoretic planning and hybrid human-AI system design (Basye et al., 1992; Dellermann et al., 2021).

Governance evaluation synthesizes recent policy and auditing work to derive practical accountability measures. We examine dataset auditing frameworks, algorithmic transparency and explainability proposals, and normative recommendations for meaningful human control (European Parliament, 2022; Rossi & Mattei, 2019; Santoni De Sio et al., 2022). Each recommendation is evaluated for feasibility and potential unintended consequences.

Finally, in constructing a layered architecture, we combine insights from the conceptual mapping, technical analysis, and governance evaluation. We produce hypothetical scenario analyses—both sacrificial trolley-like dilemmas and more prosaic collision avoidance scenarios—to illustrate how the architecture functions under uncertainty. The methodology prioritizes coherence between normative justification and mechanical feasibility, arguing that ethical correctness depends on both.

Results

This section presents the outcomes of the conceptual mapping, technical analysis, and governance evaluation and synthesizes them into the layered ethical architecture. The results are descriptive and theoretical: they identify constraints, propose mappings, and demonstrate how ethical principles can be operationalized without resorting to sensationalized sacrificial logic.

Mapping moral theories to algorithmic primitives

Utilitarianism and outcome-based objective functions. Utilitarian ethics prioritizes aggregate outcomes, suggesting that AV systems should minimize expected aggregate harm. In algorithmic practice, this equates to objective functions that weight outcomes (e.g., injury severity and casualty counts) and optimize expected value under uncertainty. Decision-theoretic planners and stochastic control frameworks naturally instantiate such objectives through expected-utility maximization (Derek, 2017; Basye et al., 1992). However, strictly utilitarian implementations face two key constraints: (1) measurement and valuation problems in quantifying harms and comparing across individuals; and (2) the computational and informational limits of real-time prediction under noisy perception (Zhang et al., 2018). Hennig & Hütter (2020) further emphasize that human moral judgments do not collapse neatly into utilitarian calculus; models of human dilemma response require nuanced parameters that can diverge from pure consequence-maximizing rules.

Deontology and rule-based constraints. Deontological ethics emphasizes duties and prohibitions that should not be violated even for beneficial outcomes. Algorithmically, this maps to hard constraints or rule-based overrides—e.g., do-not-target-pedestrians, do-not-intentionally-sacrifice-occupants—that the planner is forbidden to violate regardless of expected outcomes (Sven & Smids, 2016; Wendell & Colin, 2008). Implementing deontological constraints poses challenges when constraints conflict (e.g., a rule forbidding harm to one class but allowing unavoidable harm to another), leading to specification dilemmas and gridlock in optimization. Moreover, rigid hard constraints can reduce adaptability in unpredictable situations, potentially increasing overall harm if rules are infeasible in physical execution (Fossa, 2023).

Doctrine of double effect and intention-sensitive architectures. The doctrine of double effect separates intended harms from side-effects; in AV terms, this suggests architectures that distinguish between harm caused as a side-effect of pursuing a legitimate goal (e.g., braking to avoid a collision results in a secondary harm) and harm that is an intended outcome (e.g., actively steering into a pedestrian to avoid hitting several others). Translating this into systems design requires traceable decision rationale and action-selection processes that can be audited for intent-like structures (Philippa, 1967; Judith, 1976). Implementing such intent-sensitive features benefits from explainable planning layers that record decision contexts and counterfactual reasoning paths but is complicated by the fact that machine "intention" is not the same as human intention; the architecture must therefore operationalize intention as an artifact of goal selection and constraint satisfaction rather than phenomenological intention (Jianwu, 2018).

Hybrid approaches and ethically bounded AI. The most practical path is a hybrid architecture combining outcome-aware optimization with layered constraints and traceable decision rationale (Rossi & Mattei, 2019; Dellermann et al., 2021). This approach uses a primary objective to minimize expected harm but enforces a set of protected constraints (e.g., preserving occupant safety as a priority, respecting traffic laws) and records decision logs for ex post audits. The hybrid model recognizes moral pluralism and attempts to operationalize multiple normative considerations simultaneously.

Technical constraints shaping ethical choices

Perception uncertainty and limited situational awareness. The AV's perception module (sensor fusion and deep learning-based perception) is the source of the agent’s knowledge and therefore directly shapes ethical feasibility. Misclassifications, occlusions, and adversarial vulnerabilities can produce incorrect world models, thereby changing the set of feasible actions. Any ethical architecture must therefore be robust to perceptual error and conservative when uncertainty is high (Zhang et al., 2018; Keqiang, 2017). The necessity of conservative policies under uncertainty may preclude certain sacrificial options that depend on precise, high-confidence classifications.

Prediction uncertainty and stochasticity of agent behavior. Trajectory prediction for other road users is probabilistic; planners must therefore work over distributions of possible futures (Basye et al., 1992). This requirement complicates utilitarian calculations because expected harm depends on prediction distributions that may be multimodal and heavy-tailed. Hence, planners must employ risk-sensitive objectives (e.g., conditional value-at-risk) and robust optimization techniques to balance average-case and worst-case considerations (Basye et al., 1992).

Control and feasibility constraints. Some theoretically optimal actions may be physically infeasible due to vehicle dynamics, road geometry, or environmental conditions. The set of safe maneuvers is constrained by braking distances, steering limits, and the presence of obstacles, which implies that ethical reasoning must be embedded within physically constrained planning modules (Keqiang, 2017). This further distances real AV behavior from philosophical thought experiments that assume unconstrained, instantaneous action.

Governance and data quality implications

Dataset bias and representativeness. The data used to train perception and planning modules strongly influence the AV's behavioral tendencies. Systematic biases in datasets (e.g., underrepresenting certain pedestrian demographics or environmental conditions) can skew performance and lead to differential safety outcomes across populations (European Parliament, 2022). Auditing dataset quality and representativeness is therefore an ethical imperative (Patil et al., 2025; European Parliament, 2022).

Auditability and explainability. To operationalize moral reasoning and to enable legal and societal oversight, AVs must generate decision logs and explanations amenable to audit (Rossi & Mattei, 2019; Williams et al., 2022). Explainability in this context is not merely model interpretability for researchers; it is a structured record documenting why specific actions were chosen under specific constraints and uncertainties. This record must be standardized and protected for both transparency and privacy concerns.

Meaningful human control and human-AI collaboration. Given the limits of fully autonomous ethical reasoning, the literature recommends mechanisms for meaningful human control—clear interfaces for human intervention, policy-level oversight, and design decisions that preserve human agency in deployment contexts (Santoni De Sio et al., 2022; Dellermann et al., 2021).

Nonetheless, the feasibility of real-time human intervention in high-speed driving remains contested; thus, human control must often be interpreted at governance and design-time levels rather than as instantaneous overrides.

Synthesis: A layered ethical architecture

Based on the foregoing, we propose a layered architecture with three interacting strata:

1. Reactive Safety Layer (low-level): Real-time collision mitigation, physical constraints, and failsafe control—implemented as hard real-time controllers that enact the most conservative, physically feasible maneuvers to avoid imminent harm (Keqiang, 2017). This layer operates under high reliability and low latency and is agnostic to complex normative trade-offs; its ethical role is to minimize immediate harm consistent with vehicle dynamics and sensor confidence.

2. Decision-Theoretic Planning Layer (mid-level): A stochastic planner that optimizes over expected harm subject to protected constraints. This layer integrates prediction uncertainty, risk-sensitive objectives, and rule-based prohibitions reflecting deontological commitments or legal constraints. It provides a candidate action set for the reactive layer and records counterfactuals and rationale for auditing (Derek, 2017; Basye et al., 1992).

3. Governance and Audit Layer (high-level): Policy rules, dataset governance, model auditing, and human oversight mechanisms. This layer sets the normative boundaries (safety targets, fairness constraints, transparency requirements), mandates dataset quality standards, and ensures accountability and compliance with legal frameworks (European Parliament, 2022; Rossi & Mattei, 2019; Williams et al., 2022).

Hypothetical scenario analyses. We apply the architecture to two illustrative scenarios. First, a classic trolley-like scenario where a vehicle faces an imminent choice: swerve into one pedestrian to avoid hitting five. In realistic sensor and control constraints, the Reactive Safety Layer may preclude the lateral maneuver due to insufficient steering authority; the Decision-Theoretic Planner would therefore consider only feasible options and prioritize braking and avoidance that minimize expected harm across all agents, but would not be able to execute a sacrificial intentional steering maneuver if control constraints make it infeasible. The Governance Layer would ensure that decisions are logged and evaluated against policy. Second, a complex urban incident with occluded pedestrians and ambiguous predictions: the architecture favors conservative maneuvers, risk-sensitive planning, and possible early external communications (e.g., honking, V2X alerts) to reduce reliance on sacrificial choices.

Discussion

The previous section's synthesis yields several deep implications for ethical theory, engineering practice, and policy. We discuss normative, technical, and governance-level consequences and critically examine limitations of the proposed layered architecture as well as avenues for future work.

Normative implications: moving beyond sacrificialism. The dominance of trolley-problem narratives in public discourse has skewed attention toward edge-case sacrificial dilemmas and away from quotidian but ethically salient design choices such as dataset selection, sensor placement, and conservative policy design (Zhao & Li, 2020; Wu, 2020). Our analysis suggests that ethically significant decisions often occur upstream of last-moment collision trade-offs: in choosing what data to collect, which populations to represent, how to weight different types of harm in objective functions, and in setting safety thresholds. Philosophically, this points to a shift from judging single instances of sacrificial choice to evaluating systemic decisions that shape risk distributions across populations (Hennig & Hütter, 2020; European Parliament, 2022).

Reconciling moral pluralism in system design. Moral theorists disagree profoundly about foundational principles—utilitarians emphasize aggregate welfare, deontologists emphasize duties, and virtue ethicists emphasize character and context. The hybrid layered architecture accommodates pluralism by allowing multiple normative considerations to manifest at different levels: constraints or prohibitions at the governance layer (reflecting deontological commitments), expected-harm optimization at the decision-theoretic layer (reflecting utilitarian aims), and conservative control heuristics at the reactive layer (reflecting prudential or virtue-like caution). This pluralistic architecture recognizes that no single ethical approach suffices and that practical systems must balance competing normative demands.

Technical trade-offs and the ethics of uncertainty. The prominence of probabilistic perception and prediction

systems in the AV stack forces ethical agents to reason under uncertainty. Ethical decision-making under uncertainty must therefore consider not only expected outcomes but also distributional risk and tail behavior. For instance, choosing an action that minimizes expected harm but increases the variance of outcomes (including small but catastrophic tail events) may be normatively problematic. Risk-sensitive optimization frameworks (e.g., CVaR minimization) offer a technical path to encode ethical concerns about tail risks but require normative grounding to set acceptable risk thresholds (Basye et al., 1992). Moreover, technical mitigations such as uncertainty quantification, conservative policies under low confidence, and active information-gathering (evasive maneuvers that improve visibility) can reduce reliance on ethically fraught trade-offs.

The role of datasets and the politics of representation. Data are not ethically neutral; dataset composition and annotation schema embed value judgments about what constitutes "typical" or "acceptable" behavior and which scenarios receive attention. Auditing for demographic and environmental representativeness becomes an ethical imperative because underrepresentation leads to unequal safety outcomes (European Parliament, 2022; Patil et al., 2025). Furthermore, the procedures for labeling injury severity, vulnerability, and risk involve normative choices—are certain occupational groups weighted differently? How are children accounted for? These choices must be transparent and subject to public deliberation. A governance layer that mandates dataset auditing and public reporting can make such choices accountable.

Responsibility and legal alignment. Attributing moral and legal responsibility in AV incidents is complex because multiple stakeholders contribute: vehicle manufacturers, software suppliers, dataset curators, municipal authorities maintaining infrastructure, and end-users. The layered architecture provides a framework for attributing responsibilities by mapping decisions to layers: reactive controller failures suggest engineering or maintenance lapses; failures in planning may indicate design or specification problems; governance failures reflect systemic regulatory shortcomings. Legal systems must evolve to handle distributed liability models, possibly combining product liability, strict liability for operating entities, and regulatory penalties for governance non-compliance (Wu, 2020). Clear standards for logging and explainability are prerequisites for fair adjudication (Rossi & Mattei, 2019; Williams et al., 2022).

Meaningful human control: definitional and practical challenges. The literature advocates for meaningful human control, but defining operational criteria is nontrivial (Santoni De Sio et al., 2022). Real-time human override is infeasible at highway speeds given human reaction times; therefore, meaningful human control must be interpreted across temporal and institutional scales. Design-time control includes human-in-the-loop validation of objectives and constraints; deployment-time control includes remote oversight and fail-safe protocols; and governance-time control includes public accountability and standard-setting. The layered architecture supports these varied modalities by exposing design-time artifacts (objective functions, protected constraints) for review and by producing standardized logs for oversight.

Limitations of the architecture and open problems. While the layered framework reconciles many practical and normative demands, it has limitations. First, the translation of normative constraints into executable code requires quantification of inherently qualitative judgments (e.g., how much weight to assign to different injury severities), which may be contested and culturally contingent. Second, while the architecture prescribes logging for auditability, privacy concerns and data protection regulations restrict the granularity and retention of logs, creating tensions between accountability and privacy. Third, the governance layer presumes capable institutions with technical literacy and enforcement power; in many jurisdictions, such institutions do not yet exist. Finally, the architecture relies on accurate uncertainty quantification in perception and prediction modules—an area of active research with unresolved challenges (Zhang et al., 2018).

Future research directions. Several research programs emerge from this analysis. Empirically, research should quantify how dataset composition influences ethical outcomes across diverse contexts and develop standardized auditing protocols (European Parliament, 2022; Patil et al., 2025). Technically, work is needed on risk-sensitive planning and formal methods that guarantee safety properties under bounded uncertainty (Basye et al., 1992; Keqiang, 2017). Normatively, participatory processes should be designed to derive

societally acceptable weightings and protected constraints, ensuring democratic legitimacy. Finally, interdisciplinary pilot projects integrating technical, legal, and social-science expertise can test the layered architecture in controlled deployments, revealing practical frictions and refinements.

Conclusion

The trolley problem has been a productive catalyst for public engagement with machine ethics but is a poor proxy for the real ethical work required to design, deploy, and govern autonomous vehicles. This paper advances an alternative: a layered ethically bounded architecture that reconciles philosophical pluralism with technical feasibility and governance needs. The architecture recognizes that most ethically relevant decisions are not last-second sacrificial choices but upstream choices about data, objectives, constraints, and institutional design.

Concretely, we recommend the following actionable steps. First, design AV stacks with a conservative Reactive Safety Layer that minimizes immediate harm under high sensory uncertainty. Second, implement a Decision-Theoretic Planning Layer that optimizes expected harm subject to protected constraints reflecting democratically justified prohibitions. Third, institute robust Governance and Audit Layers that require dataset quality assurance, standardized decision logs, and transparent reporting. Fourth, create legally and institutionally grounded models for distributed responsibility that align with these technical mappings. Finally, prioritize public deliberation and participatory processes to determine the content of protected constraints and acceptable risk thresholds.

By reframing the debate away from sensationalized sacrificial scenarios and toward implementable, auditable architectures that integrate moral reasoning with engineering constraints, we can make meaningful progress toward AV systems that are not only technically competent but ethically legitimate. Such progress requires interdisciplinary collaboration, regulatory innovation, and a commitment to continuous auditing and public engagement.

References

1. Zhao, L.; Li, W. “Choose for No Choose”—Random-Selecting Option for the Trolley Problem in Autonomous Driving. In LISS2019: Proceedings of the 9th International Conference on Logistics, Informatics and Service Sciences; Springer: Singapore, 2020; pp. 665–672.
2. Wu, S.S. Autonomous vehicles, trolley problems, and the law. Ethics Inf. Technol. 2020, 22, 1–13.
3. Philippa, F. The problem of abortion and the doctrine of double effect. Oxf. Rev. 1967, 5, 5–15.
4. Judith, J.T. Killing, letting die, and the trolley problem. Monist 1976, 59, 204–217.
5. Derek, L. A Rawlsian algorithm for autonomous vehicles. Ethics Inf. Technol. 2017, 19, 107–115.
6. Gray, M. Moral machines. New Yorker. 2012, p. 24. Available online: https://www.newyorker.com/news/news-desk/moral-machines (accessed on 21 August 2024).
7. Wendell, W.; Colin, A. Moral Machines: Teaching Robots Right from Wrong, 1st ed.; Oxford University Press: New York, NY, USA, 2008.
8. Jianwu, L. Capacity Difference and Responsibility Difference: On Possibility of Driverless Vehicle as Ethical Subject. Soc. Sci. Yunnan 2018, 4, 15–20+186.
9. Sven, N.; Smids, J. The ethics of accident-algorithms for self-driving cars: An applied trolley problem? Ethical Theory Moral Pract. 2016, 19, 1275–1289.
10. Keqiang, L. Key topics and measures for perception, decision-making and control of intelligent electric vehicles. Sci. Technol. Rev. 2017, 14, 85–88.
11. Zhang, X.; Gao, H.; Zhao, J.; Zhou, M. Overview of deep learning intelligent driving methods. J. Tsinghua Univ. (Sci. Technol.) 2018, 58, 438–444.
12. Basye, K.; Dean, T.; Kirman, J.; Lejter, M. A decision-theoretic approach to planning, perception, and control. IEEE Expert 1992, 7, 58–65.
13. F. Fossa, Unavoidable Collisions. The Automation of Moral Judgment. Stud. Appl. Philos. Epistemol. Ration. Ethics, vol. 65, pp. 65–94, 2023, doi: 10.1007/978-3-031-22982-4_4/COVER.
14. Parliamentary Forum for the Study of Science and Technology. Auditing the quality of datasets used in algorithmic decision-making systems. Eur. Parliam. Res. Serv., 2022, doi: 10.2861/98930.
15. M. Hennig and M. Hütter, Revisiting the divide between deontology and utilitarianism in moral

dilemma judgment: A multinomial modeling approach. J. Pers. Soc. Psychol., vol. 118, no. 1, pp. 22–56, Jan. 2020, doi: 10.1037/PSPA0000173.

16. C. Wu; R. Zhang; R. Kotagiri; P. Bouvry, Strategic Decisions: Survey, Taxonomy, and Future Directions from Artificial Intelligence Perspective. ACM Comput. Surv., vol. 55, no. 12, Mar. 2023, doi: 10.1145/3571807.

17. D. Dellermann; A. Calma; N. Lipusch; T. Weber; S. Weigel; P. Ebel. The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. Proc. Annu. Hawaii Int. Conf. Syst. Sci., vol. 2019-January, pp. 274–283, May 2021, doi: 10.24251/hicss.2019.034.

18. Patil, A. A.; Patel, N.; Deshpande, S. Ethical Decision-Making In Sustainable Autonomous Transportation: A Comparative Study Of Rule-Based And Learning-Based Systems. International Journal of Environmental Sciences, 11(12s), 390-399, 2025.

19. F. Rossi; N. Mattei. Building Ethically Bounded AI. Proc. AAAI Conf. Artif. Intell., vol. 33, no. 01, pp. 9785–9789, Jul. 2019, doi: 10.1609/AAAI.V33I01.33019785.

20. D. Shin. User Perceptions of Algorithmic Decisions in the Personalized AI System: Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability. J. Broadcast. Electron. Media, vol. 64, no. 4, pp. 541–565, Oct. 2020, doi: 10.1080/08838151.2020.1843357.

21. R. Williams et al. From transparency to accountability of intelligent systems: Moving beyond aspirations. Data Policy, vol. 4, no. 3, p. e7, Feb. 2022, doi: 10.1017/DAP.2021.37.

22. F. Santoni De Sio; G. Mecacci; S. Calvert; Daniel Heikoop; M. Hagenzieker; B. Van Arem. Realising Meaningful Human Control Over Automated Driving Systems: A Multidisciplinary Approach. Minds Mach. 2022, pp. 1–25, Jul. 2022, doi: 10.1007/S11023-022-09608-8.

23. H. Karvonen; E. Heikkilä; M. Wahlström. Safety challenges of ai in autonomous systems design – solutions from human factors perspective emphasizing ai awareness. Lect. Notes Comput. Sci., vol. 12187 LNAI, pp. 147–160, 2020, doi: 10.1007/978-3-030-49183-3_12.