THE MORALITY OF AI AGENTS IN AUTONOMOUS VEHICLE DECISION-MAKING

Toward a Pluralistic and Adaptive Solution

Eirini Ornithopoulou

Word count: **4184** without Title page, TOC or AI Use Statement

Coursework for Philosophy of Artificial Intelligence: Ethics and Policy
HT 2024, as taught by Assistant Professor Dimitri Mollo at Umeå University

Table of Contents

Introduction

Recent advancements in AI have catalyzed fundamental changes in how we envision transportation, most notably through autonomous vehicles (AVs). Unlike human drivers, who act on personal moral intuitions and social conditioning, AVs rely on algorithmic processes to make split-second decisions that make the difference between life and death. To illustrate, consider a scenario in which an autonomous car must decide whether to swerve to avoid a pedestrian—even at the risk of harming its own passengers—or to continue straight ahead, prioritizing its passengers' safety while endangering jaywalkers. Such moral conundrums echo versions of the *trolley problem*, a thought experiment that has historically tested our intuitions about sacrificing one life to save many.

While it might be tempting to treat these AI dilemmas purely as engineering challenges, the deeper ethical stakes require robust philosophical scrutiny. AI systems, after all, lack human intentionality, emotions, or empathy. They cannot reason about right or wrong in the same way humans do; any "moral" judgment is mediated by lines of code, data-driven models, and optimization targets. Moreover, humans have broad, culturally contingent notions of what constitutes acceptable risk or moral action. When a crash becomes unavoidable, how should the AV distribute harm, and whose lives take priority?

These questions underscore a central theme of this essay: the moral programming of autonomous vehicles cannot be resolved by technical fixes or a single ethical principle. Rather, this essay argues that we must integrate *ethical pluralism*—the view that multiple valid ethical frameworks can coexist—alongside rigorous accountability structures, transparency, and adaptive learning mechanisms. By weaving together philosophical theories (from utilitarianism to deontology), cultural relativism, empirical findings (e.g., the Moral Machine Experiment (Awad et al., 2018)), and practical constraints (such as legislative gaps), this essay contends that only a multifaceted approach will ensure moral legitimacy and public trust.

As we explore these intersections, we will examine how classical philosophical problems gain renewed relevance in an era of AI-driven technologies. Although the present focus is on AVs, the discussion illuminates broader questions about responsibility, algorithmic opacity, and the alignment of AI with human values. Ultimately, the ethics of autonomous vehicles becomes a microcosm of a larger, pressing question: *How can societies guide AI in a manner that respects diverse norms while preserving human dignity and autonomy*?

1. Philosophical Background: The Trolley Problem and Moral Theories

1.1 The Emergence of the Trolley Problem

The *trolley problem* traces its origins to philosophers such as Philippa Foot (1967) and Judith Thomson (1976). Foot's original formulation describes a runaway trolley on course to kill five workers. A bystander can pull a lever, diverting the trolley onto another track where it kills only one person instead of five. Thomson added further variations, such as the "fat man" scenario, which forces one to contemplate physically pushing an individual onto the track to

halt the trolley. These scenarios test one's moral intuitions about whether—and when—it is permissible to cause harm to save more lives.

Philosophers typically use these thought experiments to illuminate tensions among major ethical theories:

- **Utilitarianism**: contends maximizing overall welfare or utility. In the trolley problem, a utilitarian might argue that saving five lives at the expense of one is morally justified because it yields a net gain in well-being.
- **Deontological Ethics**: emphasizes adherence to moral duties and rules, such as "do not intentionally kill." A strict deontologist might maintain it is morally impermissible to kill one person, even if it saves five, because doing so violates an inviolable moral rule.
- **Virtue Ethics**: focuses on the character and intentions of the moral agent. What kind of person would choose to take an action that causes a harm—even if beneficial overall?

Although it is a deliberately artificial dilemma, the trolley problem retains significant teaching power because it strips away extraneous factors and highlights our conflicting ethical instincts. When transplanted into real-world technology—such as AI in AVs—these clearcut dichotomies reveal deep complexities when we introduce cultural, social, and legal dimensions.

1.2 Why the Trolley Problem Matters for AVs

Autonomous vehicles are effectively placed in the trolley-driver seat. Rapid sensor-based decisions about braking, acceleration, and steering can result in scenarios eerily similar to the trolley problem. Should the car protect pedestrians at the expense of injuring its own passengers, or vice versa? Is there an ethically defensible rationale for any prioritization scheme?

One might argue that an AV's decisions are less about moral reasoning and more about probability and data analysis—yet that distinction only emphasizes the ethical stakes. If the algorithms embedded in an AV weigh certain outcomes more heavily (e.g., preserving passenger life), that is, in essence, making an ethical choice. Thus, moral reasoning is effectively coded into the vehicle's decision matrix, even if no explicitly human-like consciousness is involved.

From a philosophical perspective, the direct connection between trolley-like thought experiments and AV programming raises profound questions: can a machine be considered a moral agent? Are we comfortable delegating life-and-death decisions to non-human systems? How do we interpret blame or accountability when something goes wrong? These lines of inquiry form the basis of subsequent sections, as we examine how moral theories can—or cannot—be translated into code.

2. AI and AV Decision-Making: Real-World Challenges

2.1 Translating Morality into Code

Implementing moral judgments in AVs is complicated by practical and ethical constraints. Real streets are crowded with variables: weather conditions, unpredictable human actions, sensor

malfunctions, and so on. Programmers attempting to incorporate moral rules—say, a bias toward minimizing total harm—quickly confront technical limits. As Tolmeijer et al. (2024) note, AVs already struggle with basic object detection under adverse weather or lighting. Expecting them to identify each person's age or vulnerability in milliseconds before a crash is simply unrealistic.

Moreover, moral theory often deals with broad principles (e.g., "Do not kill an innocent person intentionally") that become ambiguous when operationalized in code. A strict deontological approach might require elaborate exceptions: "Swerving is permissible only if it does not constitute intentional harm to a bystander." What qualifies as "intentional" in an algorithm's perspective? In practice, developers must create heuristics, thresholds, and prioritization lists that do not map cleanly onto philosophical treatises.

2.2 Empirical Data: The Moral Machine Experiment

Despite these complexities, researchers have tried to collect data on public moral intuitions about AV behavior. The most prominent initiative, the *Moral Machine Experiment* (MME) by Awad et al. (2018), collected over 40 million decisions from participants worldwide. Respondents were presented with hypothetical crash scenarios akin to the trolley problem—saving pedestrians vs. passengers, young vs. old, humans vs. animals—and asked which outcome they deemed preferable.

Results revealed substantial cultural variation. For instance, "Western" participants often leaned toward saving the greater number of people (a more utilitarian approach), while some "Eastern" regions showed a modest bias against making active interventions that harm others, echoing a more deontologically favored outlook. Other participants displayed preferences seemingly shaped by local norms, such as sparing the young over the elderly or prioritizing law-abiding pedestrians over jaywalkers.

Critics of the MME point out that forced binary choices can distort participants' true moral intuitions, and real-life accidents rarely unfold with such clarity. Nonetheless, these findings highlight *cultural relativism* in moral reasoning, reinforcing the difficulty of crafting a single universal moral code for AVs. As Bigman and Gray (2020) observe, moral dilemmas only scratch the surface of the broader sociological and psychological factors that shape public acceptance of AV technologies.

2.3 Legal and Regulatory Barriers

Beyond theoretical or empirical questions, designers of AV moral systems must navigate legal frameworks that do not easily accommodate machine-based decisions. In many jurisdictions, laws assume a human driver's intentionality and capacity for negligence. If an AV is at fault in an unavoidable crash, should liability rest on the manufacturer, the software developer, the vehicle owner, or even the AI system itself? Wu (2019, 2020) underscores that ambiguous legal settings disincentivize robust moral programming: manufacturers may fear lawsuits if they encode certain risk choices that later prove controversial.

Such regulatory uncertainties often encourage minimal moral programming—aiming to meet standard road safety guidelines without delving into ethically fraught territory. Yet, as AVs become increasingly sophisticated, accidents may be less about oversight and more about moral

prioritization: the rare but inevitable scenario where the vehicle must choose a lesser evil. This friction between engineering constraints and legal liability further justifies the need for an overarching philosophical framework that can guide regulators, developers, and the public.

---

3. Cultural Relativism and Ethical Pluralism

3.1 The Limits of a Universal Moral Code

The notion of *ethical pluralism* counters the assumption that a single moral framework can be seamlessly applied across all contexts. The Moral Machine Experiment's cross-cultural findings (Awad et al., 2018) illustrate how even seemingly fundamental moral questions—like whether to save the many over the few—can provoke starkly different answers. Factors such as religious tradition, shared values, and historical experiences shape how societies conceive of moral responsibility.

In practical terms, an AV in one region might best align with local norms by emphasizing collective welfare—i.e., saving the majority at all costs. Meanwhile, another region might bristle at the idea of intentionally steering harm toward the few, even if it saves more lives. At a policy level, expecting one universal algorithm to satisfy these divergent intuitions is unrealistic. Ethicists of a communitarian bent would go further, arguing that moral codes arise from community relationships and cannot be simply transplanted into a foreign environment.

3.2 Ethical Pluralism in Action

Ethical pluralism posits that multiple moral viewpoints can hold legitimacy simultaneously. Rather than labeling them strictly right or wrong, pluralism suggests that each perspective might be suitable depending on context, cultural background, or specific scenario details. For instance, a deontological principle ("never deliberately harm an innocent person") could be combined with a utilitarian sensitivity to overall harm reduction.

In the domain of autonomous vehicles, ethical pluralism might manifest in adaptive algorithms that weigh different normative considerations dynamically. Under typical road conditions, the AV might preserve passenger safety, reflecting the usual understanding of a vehicle's role. In more extreme situations (e.g., a crowded crosswalk with young children), the system might temporarily prioritize a utilitarian logic of minimizing total harm. Such a hybrid approach tries to integrate the strengths of each theory while acknowledging that moral life is rarely black and white.

However, implementing ethical pluralism is no simple feat: it requires reconciling potentially conflicting moral claims, setting up robust guidelines for when one principle takes precedence, and deciding whose cultural norms shape these preferences. If developers allow local governments or community boards to calibrate these "ethical dials," we would be fragmenting moral standards across jurisdictions. Conversely, a top-down universal approach overlooks vital cultural nuances. These debates resonate in real-world AI governance, underscoring why moral alignment is more than a technical puzzle.

4. Accountability and Transparency

4.1 Who is to Blame?

One of the thorniest moral issues with AVs is the question of accountability. Unlike human drivers, AI systems do not have consciousness, intentionality, or a sense of guilt. When something goes tragically wrong, we cannot simply say "the AI is morally responsible." Instead, blame must be attributed among various actors:

- Developers who coded the decision-making algorithms,
- Manufacturers who integrated these algorithms into vehicles,
- Regulatory Bodies that set standards, or
- Vehicle Owners who deployed these systems on public roads.

This distribution of responsibility becomes murky if an AV makes a decision that aligns with a pre-set ethical rule (e.g., prioritizing the passenger) but that society later deems unacceptable. According to Vaassen (2022), there is an inherent risk that no one will step forward to accept responsibility because each stakeholder can claim it was simply following instructions, designs, or legal requirements.

4.2 Opacity and the Black Box Problem

Accountability is intertwined with *algorithmic transparency*. Modern AI, especially machine learning and deep neural networks, often yields "black-box" models whose internal reasoning is difficult to interpret, even for their creators. It might be unclear why the system decided to swerve left rather than right. This lack of interpretability erodes personal autonomy and the capacity of external observers—whether victims or the court—to contest or understand the moral logic behind the outcome.

Philosophically, opaque decision-making processes undermine core principles of justice. Without a clear explanation for why a car took a certain action, how can a legal body assess culpability? Vaassen (2022) further argues that such opacity threatens the Kantian ideal of treating individuals as ends rather than means, because they effectively become passive subjects of decisions they cannot question or influence.

4.3 Approaches to Building Trust

To mitigate these dangers, scholars increasingly advocate for *Explainable AI* (XAI) in safety-critical domains (Machado et al., 2024). XAI approaches strive to generate human-interpretable rationales for AI decisions, whether through simpler models (like decision trees) or techniques that highlight which factors were most influential at each step. While some critics worry that fully interpretable models might reduce performance, the ethical imperative for accountability in scenarios with life-and-death stakes is often deemed to override minor performance gains.

In addition to technical transparency, trust can also be fostered through participatory governance. Communities could be given a voice in shaping the moral parameters of local AV deployment. Through public consultations, panel discussions, and ballot votes, stakeholders learn about the trade-offs and help define the baseline values. Such an approach does not

guarantee universal consensus, but it at least distributes the moral burden and fosters informed dialogue.

5. Proposed Framework: A Hybrid Pluralistic Model

5.1 Moral Proxy and the Limits of Static Codes

The notion of a "moral proxy" suggests that AI systems should directly encode a moral framework chosen by human developers (Evans et al., 2020). In essence, the AI would function as an extension of its creators' ethical reasoning. For instance, engineers might adopt a utilitarian code: "Always minimize total harm." However, critics note that this approach rigidly ties the AI to a single, potentially controversial standpoint. If public sentiment later shifts or if local cultural values differ sharply, the system's moral logic risks appearing tyrannical or misaligned with those it serves.

Furthermore, the moral proxy concept raises the question: *Whose moral judgments are being proxied*? The developer's personal beliefs? A committee's consensus? An international standard? In a world of plural moral systems, no single perspective can claim universal authority—especially when these vehicles cross national borders, each with distinct cultural mores.

5.2 Integrating Adaptive Learning

To address the dynamic and context-dependent nature of ethics, many scholars propose that AVs use *adaptive learning* mechanisms capable of refining moral priorities over time (Machado et al., 2024). Through real-world feedback—such as user input, accident data, or updated social norms—an AV's moral algorithms could recalibrate. For instance, if local communities voice strong objections to saving the passenger over multiple pedestrians, the AV might gradually shift the weighting in its risk calculus.

While appealing in theory, adaptive learning also invites potential pitfalls. Continuous changes in moral logic could sow confusion about which guidelines were active at the moment of a crash. If moral parameters shift too frequently or unpredictably, holding stakeholders accountable becomes even harder, as it may be unclear which version of the algorithm was responsible. Likewise, critics warn against opening moral protocols to bias infiltration: if certain vocal groups can push updates that privilege their interests, minority rights or broader notions of justice might be overshadowed (Bigman & Gray, 2020).

5.3 Ethical Pluralism as a Mediating Framework

The hybrid approach proposed here incorporates both static and adaptive elements within a framework of *ethical pluralism*. Broadly, the system would:

1. Start with a Core Set of Universal Protections – Perhaps drawn from widely recognized principles, such as avoiding intentional harm or discriminating based on arbitrary traits (e.g., race, religion).

2. Allow Context-Specific Calibration – Developers or local governance bodies may set "weighting" parameters that reflect local norms—whether an emphasis on majority welfare or on absolute respect for individual rights.
3. Implement Iterative Feedback Loops – Over time, the system adjusts or refines these weightings based on real-world performance, accident data, and shifts in community sentiment.
4. Maintain Transparency and Accountability Mechanisms – Through Explainable AI techniques and mandatory logs, each moral decision or adjustment remains open to inspection by regulatory bodies, ethicists, and the public.

By preserving a baseline moral foundation while enabling iterative learning, AVs can adopt a philosophical stance that respects the multiplicity of moral perspectives yet avoids the extremes of moral relativism. The overall goal is *contextual stability*, wherein the system can handle exceptional cases without losing all sense of consistent responsibility.

5.4 A Philosophically Informed Governance Model

This essay's proposition is less about prescribing a particular moral outcome and more about outlining a *governance structure* grounded in philosophy, law, and participatory input. Specifically:

- Multidisciplinary Ethical Audits: Panels of ethicists, engineers, psychologists, and legal scholars should periodically review the AV's moral decision framework, as well as use cases. This group could also incorporate community representatives, ensuring that local sensibilities influence policy.
- Publicly Accessible Ethical Parameters: To enhance trust, key moral settings—like passenger vs. pedestrian bias—would be disclosed in a manner accessible to non-experts.
- Adaptive but Bound: While the algorithm can learn from real-world outcomes, it remains bound by certain non-negotiable constraints (e.g., not intentionally targeting vulnerable individuals or groups, preserving basic rights).
- Long-Term Deliberation: As both technology and societal norms evolve, so too should the guidelines. This demands a formal process for re-assessing the moral framework, ensuring that it remains democratically legitimate.

By balancing the desire for algorithmic adaptation with the need for stable moral anchors and open accountability, this model aspires to embed a spirit of continuous ethical reflection within AI design. It addresses the alignment problem not as a one-off coding challenge but as an ongoing philosophical project that demands rigorous scrutiny over time.

6. Emotions, Moral Deskilling, and Human Agency

6.1 The Role of Emotions in Moral Judgment

Human moral judgment is not purely rational. Emotions such as empathy, fear, guilt, and compassion often drive decision-making more strongly than cold calculations. One might argue that removing these emotional nuances from an AV's moral logic leads to outcomes that feel *intuitively* unethical, even if they pass a rational cost–benefit test. Bigman and Gray (2020) discuss how moral decisions that appear "logically correct" can arouse public horror if they

lack any semblance of empathy—imagine an AV that unemotionally runs over a senior citizen because a purely utilitarian principle deemed it optimal.

Philosophically, the question arises: can or should AI emulate emotional reasoning? Some view emotions as essential for genuine moral agency, while others see them as evolutionary heuristics that can lead to bias or error. In the context of AVs, fully replicating human emotionality is neither technically feasible nor necessarily desirable. Still, systems designed without consideration of emotional resonance risk losing public trust. A bridging approach may involve building in constraints that mirror typical human moral sentiments, like a "do not harm children" principle, even if the rational calculus would otherwise differ.

6.2 Moral Deskilling and the Erosion of Agency

A related concern is "moral deskilling," the notion that by outsourcing moral judgments to machines, humans may gradually lose their capacity to navigate ethical complexity (Himmelreich, 2018). If we grow accustomed to AVs making all the tough decisions, perhaps society becomes less critical or reflective about moral dilemmas in general. Overreliance on technology could lead to complacency, dampening the sense of personal responsibility vital for ethical growth.

Opponents of this view might argue that AVs reduce accidents and thereby *improve* public safety, freeing humans from the burden of making (and making bad judgements in) high-stakes situations under pressure. Still, from a virtue-ethical standpoint, repeated moral engagement is part of developing good character. Delegating too many moral decisions to AI might undermine the cultivation of virtues like courage or prudence. Balancing efficiency gains with moral agency thus forms part of a broader philosophical debate about how far humans should cede control to autonomous systems.

6.3 Implications for Human Autonomy

If accountability and transparency are insufficient, individuals might find themselves at the mercy of opaque AI. Especially when AI extends beyond vehicles to domains like hiring, policing, or healthcare, the risk grows that people become "moral bystanders," subject to machine decisions they cannot influence. Vaassen (2022) interprets this as a loss of autonomy, a condition where individuals are denied the capacity to shape the moral order of their lives.

Autonomous vehicles, in that sense, are not just about transportation. They foreshadow a future where AI-mediated decisions influence daily life. Hence, the philosophical conversation about "how an AI should act" fuses with "how humans preserve agency and hold technology accountable."

7. Conclusion

7.1 Reconciling Morality and Machine Intelligence

As society moves closer to widespread deployment of autonomous vehicles, ethical questions once confined to thought experiments have become very real. From the vantage point of the *trolley problem*, we see how deeply contested our moral intuitions can be—both across cultures

and within the same community. Despite the persistent allure of a single ethical formula (like "save the most lives"), real-world contexts, cultural heterogeneity, and legal constraints demand a richer, more flexible framework.

7.2 Central Insights

1. No Single Ethical Theory Suffices: Utilitarian approaches might maximize overall benefits but alienate those who view certain harms as categorically impermissible. Deontological stances protect rights but can yield outcomes that seem socially suboptimal in extreme cases. Virtue ethics underscores moral character but offers limited guidance for machine-coded rules.
2. Cultural Relativism Is Real: Empirical evidence from the Moral Machine Experiment (Awad et al., 2018) shows how moral intuitions differ across societies, undermining the notion of a uniform moral algorithm.
3. Accountability and Transparency Are Key: Opacity in AI systems undermines trust and moral agency, calling for explainable AI solutions and robust regulatory oversight. Legal structures must adapt to distributed, machine-involved responsibility.
4. Ethical Pluralism Serves as a Path Forward: By combining multiple moral principles, AV decision systems can balance universal protections with local cultural adaptations. Adaptive learning can fine-tune these principles, though safeguards remain necessary to prevent bias and confusion.

7.3 Toward a Pluralistic and Adaptive Moral Alignment

This project work has proposed a *hybrid, pluralistic model* wherein autonomous vehicles are guided by a layered ethical framework. A baseline set of universal protections ensures some moral consistency, while context-specific parameters reflect local values. Iterative feedback loops enable adaptation over time, yet remain bounded by accountability mechanisms, multidisciplinary audits, and public oversight.

In essence, developing moral AVs is not merely a technical puzzle. It demands collaboration between philosophers, engineers, legislators, community representatives, and indeed the broader public whose lives will be affected by AI decisions. Only through a participatory approach can we secure both moral legitimacy and social acceptance for systems that stand to profoundly shape our future.

7.4 Beyond AVs

While we have anchored the discussion on the case of autonomous vehicles, the broader repercussions touch on any domain where AI wields consequential influence: healthcare triage, military drones, judicial risk assessments, and more. In each of these areas, we must navigate tensions between universal ethics and cultural specificity, accountability and opacity, efficiency and moral agency. By engaging deeply with these challenges in the AV context, society can pioneer frameworks and best practices that inform other emerging AI applications.

Ultimately, the moral complexity of autonomous vehicle decision-making offers a uniquely urgent vantage point on an important question: *How do we guide powerful, non-human systems in ways that reflect the best of our moral aspirations, rather than the worst of our biases or*

limitations? Charting a path forward requires humility, creativity, and ethical inquiry—values that are the essence of philosophy.

References

1. **Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018).** The Moral Machine Experiment. *Nature*, 563, 59–64. https://doi.org/10.1038/s41586-018-0637-6
2. **Bigman, Y. E., & Gray, K. (2020).** Life and death decisions of autonomous vehicles. *Nature*, 579(7797), E1–E2. https://doi.org/10.1038/s41586-020-1987-4
3. **Evans, N., Tolmeijer, S., & Ju, W. (2020).** Ethical valence in autonomous vehicle decision-making: Balancing competing moral claims. *Philosophy & Technology*, 33(3), 449–464. https://doi.org/10.1007/s13347-020-00422-8
4. **Foot, P. (1967).** The problem of abortion and the doctrine of double effect. *Oxford Review*, 5, 5–15.
5. **Himmelreich, J. (2018).** Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, 21(3), 669–684.
6. **Machado, F., Vaassen, M., & Himmelreich, J. (2024).** Toward transparent and accountable AI systems: Challenges and solutions. *Journal of AI Ethics and Policy*, 12(1), 23–35.
7. **Millière, R. (2023).** The alignment problem in context: Ethical implications of in-context learning. *AI & Society*, 38(4), 701–713. https://doi.org/10.1007/s00146-023-01644-x
8. **Thomson, J. (1976).** Killing, letting die, and the trolley problem. *The Monist*, 59(2), 204–217.
9. **Tolmeijer, S., Evans, N., & Ju, W. (2024).** Navigating ethical complexity in AI-driven vehicles. *Autonomous Systems & Society*, 6(2), 101–113.
10. **Vaassen, M. (2022).** AI, opacity, and personal autonomy: Ethical considerations for explainable AI. *Philosophy and Technology*, 35(3), 489–504. https://doi.org/10.1007/s13347-022-00504-y
11. **Wu, T. (2019).** Liability and regulation in the age of autonomous vehicles. *Journal of Technology Law*, 45(2), 211–230.
12. **Wu, T. (2020).** Rethinking negligence for autonomous cars. *AI & Law Review*, 32(1), 67–85.

AI Use Statement

This philosophical essay was written with assistance from ChatGPT to enhance clarity, structure argumentation and coherence. ChatGPT was used to help clarify philosophical terms, summarize key ideas from academic papers, rephrase sentences for improved readability, and expand on initial ideas to better articulate my arguments. All content and arguments presented are my own, with ChatGPT serving as a tool to refine expression and support the writing.