Bayesian Game-Theoretic Modeling of Pedestrian-Autonomous Vehicle Interaction

Dae-Hyun Yoo[1][0009−0002−8977−1305]⋆ and
Marija Slavkovik[2][0000−0003−2548−8623]

[1] University of Pisa, Pisa, Italy
daehyun.yoo@ec.unipi.it
https://sites.google.com/view/dyoo/home
[2] University of Bergen, Bergen, Norway
marija.slavkovik@uib.no
https://slavkovik.com/

Abstract. Autonomous vehicle (AV) ethics typically positions humans as moral advisors and assumes that ethical dilemmas arise from AV malfunctions. In contrast, this article considers scenarios where immoral human intentions create ethical challenges for AVs. We propose a novel reasoning approach in AV decision-making and introduce two distinct types of AVs - outsider protection priority AVs and insider protection priority AVs - accommodating heterogeneous moral preferences among traffic participants. A static Bayesian game model is used to analyze strategic interaction between pedestrians and AVs. Our results show that not only does the presence of two distinct AV types matter, but maintaining a certain proportion of 'outsider protection priority' AVs prevents pedestrians from exploiting predictable AV features, thereby improving transportation efficiency in mixed traffic environments. Our work aims to develop practical and effective mechanisms to foster smooth and cooperative interactions between humans and AI. We hope our proposed reasoning framework encourages manufacturers to openly advertise their operating ethical principles and enhance transparency. This article contributes to the ethics of human-AI interactions, specifically in the underexplored area of conflicting moral values among different participants.

Keywords: Bayesian game theory · Intent-aware autonomous systems · Autonomous vehicle ethics · Pedestrian-autonomous vehicle interaction.

1 Introduction

Autonomous vehicles (AVs henceforth)[3] are expected to improve road safety, traffic efficiency, and fuel use [1, 2]. Yet, deploying AVs in real-world environments remains ethically and practically challenging. Roads are shared by heterogeneous

⋆ Corresponding author

[3] Autonomous vehicles are also referred to as "self-driving" or "driverless" cars; we use these terms interchangeably throughout this paper.

users - drivers, pedestrians, cyclists, and motorcyclists - whose unpredictable behavior raises difficult coordination and accountability problems [3]. A core ethical challenge in AV design concerns how these systems should be programmed to respond in dilemma situations. This issue is illustrated using variants of the classic 'trolly problem,' which explore scenarios involving life-and-death tradeoff – such as choosing between the safety of pedestrians and passengers. Such ethical dilemmas have become a central reference point in discussions of AV ethics [4–8]. Unlike human drivers, whose decisions are made in real time, AVs follow predetermined decision-making protocols. Millar [8] argues that determining how AV should respond in unavoidable crash scenarios is one of the central ethical challenges in AV development. Lin [9] emphasizes that conflicting values further complicate ethical decision-making. Although we may be reluctant to confront uncomfortable moral trade-offs, Lin contends that programmers may be required to make such decisions. They must instruct AVs on how to behave in foreseeable scenarios and embed guiding principles for responding to unforeseen ones. This demands that programmers actively engage with moral dilemmas that human drivers typically make instinctively.

Ethical challenges in AV design also extend to human-AI interactions, particularly how people anticipate, interpret, and respond to AI-driven systems [10]. March [11] reviews experimental findings from human-AI strategic interaction studies, showing that humans often adapt to computer agents and becoming more rational or self-interested. Sophisticated AI may foster such cooperation and improve outcomes in social dilemmas, reinforcing that dual potential of AVs as both targets of exploitation and as enablers of efficiency. In pedestrian-AV context, this suggest that pedestrians may learn to exploit AV's predictable caution or, conversely, engage in cooperative behavior. Reflecting this, current policy trends - such as those outlined by the Organisation for Economic Co-operation and Development (OECD) [2] - stress the need for ethical, legal, and institutional frameworks that support cooperative human-AI interaction.

This paper addresses these challenges by examining how AVs can reconcile heterogeneous moral preferences among diverse traffic participants in mixed environments. **Section 2** reviews the core ethical and coordination challenges in AV design. **Section 3** introduce the conceptual framework motivating two AV types. **Section 4** develops a game-theoretic model, considering both complete and incomplete information settings. **Section 5** presents the analytical results, and **Section 6** concludes with a discussion of design and deployment implications.

2 The Challenges of Moral Algorithms in AVs

This section introduces the ethical dilemmas associated with AVs, the heterogeneity of moral preferences among traffic participants, and the risks of programming AVs to rigidly follow rules. These challenges motivate the core problem addressed in this paper and justify the definitions used for our proposed approach and modeling.

2.1 The Ethical Dilemma and Its Importance

This paper considers moral dilemmas that arise when AVs must make decisions under conflicting ethical values.

Numerous studies emphasize the complexity of resolving competing moral preferences in AV design [4, 5, 12]. While frameworks in AI and robotics ethics exist [9, 13–16], the practical challenge lies in how to reconcile conflicting principles in machine decision-making [17]. One common approach is to embed acceptable behavior into AVs via constraints that guide decision-making in dilemma scenarios [18, 19]. A standard example is the decision of whom or what to crash into when a collision is unavoidable [20]. These cases involve moral regret: all available outcomes cause harm and no outcome is withouot cost.

Kirkpatrick [21] defines a moral dilemma as a situation in which every option violates some ethical principle. Vamplew et al. [17] add that even if AVs do not explicitly reason about ethics, their actions often imply trade-offs between protecting passengers and pedestrians. As Whittlestone et al. [12] emphasize, such trade-offs are inevitable. When values conflict, achieving more of one requires sacrificing another. Thus, effective AI ethics must move beyond abstract principles and focus on identifying and resolving tensions in concrete cases. The German Ethics Commission on Automated and Connected Driving similarly defines a dilemma as a situation involving a choice between two evils, where no clearly justifiable trade-off exists [p. 551, [22]].

In line with this literature, we define a dilemma as *a situation in which a decision must be made between two conflicting options in un unavoidable situations, where any choice results in certain harm.*

2.2 The Heterogeneity of Moral Preferences

Different stakeholders reveal varying moral preferences regarding whom AVs should prioritize in moral dilemmas [23, 24]. While scholars debate whether AVs should follow utilitarianism, deontology, or virtue ethics, there is no consensus on which principles should be encoded into AV decision-making systems [9, 14, 15, 25]. The implementation of ethics in AVs thus faces a core problem: moral pluralism.

Altogether, these findings point to a fundamental heterogeneity in AV ethics: individuals want *others'* AVs to behave altruistically but prefer *their own* to act self-protectively. Preferences shift based on perspective, cultural norms, emotional context, and expertise. Such variability challenges the feasibility of implementing a single normative framework. Instead, AV developers and policymakers may need to adopt **pluralistic strategies** - design approach that can accommodate conflicting preferences within mixed traffic environments.

2.3 The Default AV Problem

As discussed above, programing AVs to make autonomous ethical decisions is a complex and challenging task. However, most people tend to avoid engaging

with such complex moral decision-making when they must make more than a few choices. Studies across human behavioral domains - from retirement contributions to organ donations - reveal that most individuals will accept the default setting, even when alternatives are straightforward and simple [26].

In this paper, we define the default autonomous vehicle (DAV henceforth) as a highly risk-averse (or very 'defensive') and rule-following AV programmed to strictly adhere to traffic regulations under all conditions. For example, a DAV will slow down and yield to pedestrians even at unmarked crosswalks, or in cases of jaywalking, without evaluating situational context. This design, while promoting safety, introduces two major problems. First, DAVs may cause moral hazard among human road users. Because DAV behavior is deterministic and conservative, pedestrians and human drivers may begin to exploit its predictability. Second, DAVs cannot accommodate the heterogeneous moral preferences of traffic participants. Because DAVs operate under a single, uniform behavioral rule across all contexts, they fail to account for the diversity of stakeholder values in mixed traffic environments. As a result, DAVs cannot bridge the moral gap between public ethical expectations and individual self-interest - an issue that has been widely documented in AV ethics research.

3 Two Types of Moral Reasoners in AV Decision-Making

Researchers such as Bonnefon et al. [4] and Zhu et al. [27] acknowledge that AVs may be designed with divergent moral priorities. These perspectives point toward a critical distinction: AVs could operate according to distinct types of reasoners - one that prioritizes the safety of its own passengers, and one that prioritizes the protection of external road users. In what follows, we define this distinction more precisely and explain its implications for AV ethics and design.

3.1 A New Two Types of Reasoners

How should an AV reason in environments where human actors intentionally behave unethically? Unethical intentions can come from two groups of humans: those inside the vehicle and those outside it. For example, in a carjacking scenario, the AV should prioritize protection its passengers from harm. By comparison, in a terrorist attack where a vehicle is used to target pedestrians, the AV should prioritize external road users. These examples show how divergent intentions can demand different ethical reasoning. This perspective is supported by Millard-Ball [28], who warns that ill-intentioned pedestrians or drivers may exploit the predictable, risk-averse behavior of AVs.

3.2 Insiders- vs. Outsiders-Protection Priority AVs

We propose that AVs can be designed to prioritize either insiders (passengers) or outsiders (pedestrians, cyclists, and motorcyclists)[4]. This categorization is both

[4] Drivers of human-operated vehicles can also be considered outsiders relative to AVs. In some regions, cyclists and motorcyclists even outnumber pedestrians. For example,

ethically meaningful and technically implementable in real-world traffic environments. Importantly, it accommodates the heterogeneous moral preferences of traffic participants in mixed-traffic scenarios. This distinction enables manufacturers to transparently develop two AV types - outsider-protection priority AVs (**AVI** henceforth) and insider-protection priority AVs (**AVII** henceforth) - without provoking the public backlash often associated with purely self-protective AVs. This framework moves beyond traditional ethical debates in terms of utilitarianism versus deontology, by offering a practical pluralism of moral reasoning. Our framework supports the ethical and social feasibility of developing and deploying two coexisting AV types, each grounded in different moral preferences - one oriented toward internal occupants, and one toward vulnerable road users. This foundation justifies the game-theoretic analysis that follows.

4 A Game-Theoretic Approach to Autonomous Vehicles

This section introduces a game-theoretic framework to analyze strategic interactions between pedestrians and AVs. We model a scenario in which a pedestrian illegally crosses the street, intercepting the path of an AV. Before presenting our formal analysis, we first review relevant research that connects game theory with ethical reasoning and AV behavior modeling.

4.1 Related Work

The idea of using game theory to inform ethical decision-making dates to Braithwaite [29], who first argued that moral choices can be modeled through rational agent interactions. More recently, Conitzer et al. [30] propose a set of moral solution concepts in game theory, illustrating how game-theoretic tools can formalize ethical dilemmas, such as trust games, by incorporating an agent's concern for others' welfare. They suggest using extensive-form representations with passive actions as a foundation for modeling dilemmas. Game theory, in this view, offers an abstract structure for modeling moral reasoning in AI, while machine learning can be used to train agents based on these frameworks [31].

A more applied game-theoretic approach is introduced by Millard-Ball [28], who models pedestrian-vehicle interactions at crosswalks as a classic chicken game. In scenarios involving human drivers, pedestrians and vehicles respond to one another strategically, often resulting in a stable equilibrium. However, when the interaction involves AVs, assumed to be risk-averse and rule-following, pedestrians quickly learn that AVs will always yield. The more cautiously an AV behaves, the more reckless human road users may act. For example, pedestrians may step into the road without caution, or drivers may ignore traffic signals. This behavioral shift makes *Go* the pedestrian's dominant strategy, and *Stop*

in many Dutch cities, cyclists surpass pedestrians in daily traffic, while in certain parts of Italy, motorcyclists exceed pedestrians. These patterns support the inclusion of cyclists and motorcyclists as ethically relevant outsider groups in the analysis of AV decision-making.

the AV's best response - an outcome Millard-Ball [28] describes as pedestrian supremacy.[5] This outcome encourages reckless behavior, increases traffic inefficiency, and diminishes the consumer appeal of AVs that are perceived as overly submissive.

de Melo et al. [32] explore how the ethical behavior of AVs can be shaped through pre-programming by users. Their experiments show that individuals are more likely to make cooperative and fair decisions when programming AV behavior in advance, as opposed to making decisions in real-time. This suggests that autonomous systems may promote more socially cooperative outcomes if designed with precommitment mechanisms, thus offering an opportunity to foster ethical behavior at scale. Di et al. [33] introduce a hierarchical game-theoretic model to examine how optimal liability rules evolve in a mixed environment of AVs and human drivers. Their framework distinguishes strategic levels among lawmakers (top), AV manufacturers (middle), and human drivers (bottom), applying Stackelberg games to capture decision-making hierarchies. Crucially, AVs are modeled as non-strategic agents, with behavior determined by manufacturers. They argue that poorly designed liability rules may encourage moral hazard among human drivers who perceive their environment as safer due to AV presence. An optimal regulatory framework is therefore essential to balancing safety and accountability across agents.

Together, these studies illustrate the growing relevance of game theory for modeling both the strategic dimensions of AV interactions and the ethical dilemmas that arise in human-AI systems. Our work builds on this foundation by introducing a novel dual AV type model that incorporates heterogeneous moral preferences and explores how different proportions of AV types affect pedestrian behavior and transportation efficiency.

4.2 Structure of the Game

A strategic game is a model of interacting decision-makers or players, each of whom chooses from a set of actions and has preferences over the action profiles. A strategic game with cardinal preferences is defined as follows:

Definition 1 (Strategic Game with Cardinal Preferences). *A strategic game consists of:*

- *A set of players*
- *A set of actions for each player*
- *A set of cardinal preferences (payoff functions) for each player over the action profiles*

[5] 'Pedestrian supremacy' refers to a situation in which AVs behave so predictably and submissively that pedestrians exploit their risk aversion. In Section 4.3, we formalize this concept by modeling the default AV's behavior as leading to a pure Nash equilibrium in which the pedestrian always goes and the AV always stops.

Assumptions A realistic traffic situation is considered in which a pedestrian illegally crosses a street and intercepts the path of an AV. The players in this game are pedestrians and two distinct types of AVs. The two types of AVs are:

- **AVI (Type I)**: An AV that prioritizes the safety of outsiders (e.g., pedestrians, cyclists, motorcyclists)
- **AVII (Type II)**: An AV that prioritizes the safety of insiders (vehicle occupants)

Both AV types are designed to fulfill common optimization goals as acceptable common conclusions[6], such as transporting passengers safely and efficiently, minimizing fuel consumption and travel time [34]. The key ethical distinction lies in their relative prioritization of safety outcomes when moral trade-offs arise.

Each player's preferences are encoded as cardinal payoffs, which not only express the order of preference but also the degree of preference intensity. This allows us to represent ethical values and protection priorities numerically. For example, the same base payoff structure can be adjusted by attaching cardinal moral weights that reflect whether the AV prioritizes insiders or outsiders safety. In ethical decision-making under uncertainty, especially when trade-offs between safety and efficiency involved, cardinal measures are essential [39]. Using this structure, we formalize the difference between AVI and AVII as a difference in the moral weighting of potential outcomes, thereby constructing a game that captures both strategic interaction and embedded ethical reasoning.

We incorporate risk attitudes into the payoff structures for pedestrians and the two types of AVs. AVI (outsider protection priority AVs) is modeled as highly risk-averse toward pedestrians, reflecting a strong commitment to protecting vulnerable road users, such as pedestrians, cyclists, and motorcyclists. In contrast, AVII (insider protection priority AVs) is more assertive and risk-neutral toward pedestrians, as its programming prioritizes the safety of vehicle occupants over that of external parties. Pedestrians are assumed to behave risk-neutrally toward AVI, recognizing that these AVs are more likely to yield. Conversely, they act risk-aversely toward AVII, anticipating that such AVs are less accommodating and therefore pose a higher personal risk in crossing scenarios. It reflects how pedestrians anticipate more danger from assertive AVs (AVII). These risk attitudes are summarized in Table 1.

Both the pedestrian and the two types of AVs choose between two available actions: *Go* and *Stop*. All players are assumed to be rational decision-makers, choosing the best action to maximize their payoffs based on their preferences.

The interactions modeled in **Section 4.3, 4.4, and 4.5** are static games: each player chooses their action simultaneously, without knowledge of the other's choice, and the game is played only once. It is also important to note that AVs

[6] According to the floating conclusions framework, acceptable common conclusions from conflicting propositions can be justified [35, 36]. Yoo [37] applies this framework to manage heterogeneous moral preferences in ethical dilemma situations, offering support for the design of two distinct AV types: dual-AV systems.

Table 1. Risk Attitudes of Pedestrians and AV Types

	AVI	**AVII**
Pedestrian	Risk-averse / Risk-neutral	Risk-neutral / Risk-averse

Note. See Yoo [38], Chapter 2. Reproduced without revision.

in this framework are non-learning agents[7] such that they do not update their behavior or strategies based on prior outcomes or repeated interactions.

General Payoff Matrix Table 2[8] presents the general form of the payoff matrix in the interaction between a pedestrian and an AV. Rows represent the pedestrian's action choices (*Go* or *Stop*) and their corresponding payoffs; columns represent the AV's action choices (*Go* or *Stop*) and their corresponding payoffs. Each cell contains an ordered pair: the first component (e.g., a, c, e, g) denotes the pedestrian's payoff, and the second component (e.g., b, d, f, h) denotes the AV's payoff.

Table 2. Payoff Matrix of Pedestrian–AV Interaction

		AV	
		Go	**Stop**
Pedestrian	**Go**	a, b	c, d
	Stop	e, f	g, h

For the pedestrian, the highest payoff, c occurs when the AV yields to her movement. The worst-case scenario - a potential collision - is represented by the lowest payoff, a. When the pedestrian chooses to stop and yield to AV, she receives the same payoff regardless of whether the AV proceeds or stops, represented by $e = g$. The pedestrian's preference order for outcome is thus defined as:

$$c > e = g > a \tag{1}$$

This condition applies to interactions with both AVI and AVII. However, in the case of DAV, which is programmed to always yield to pedestrians regardless of

[7] As an implicit ethical agent, the AV is unable to learn or update ethical principles in response to experience, unlike a human pedestrian. This static ethical behavior aligns with the structure of a one-shot and simultaneous-move game.

[8] This table was first presented in Yoo [38]. *The Ethics of Artificial Intelligence from an Economics Perspective: Logical, theoretical, and legal discussions in autonomous vehicle dilemma*. Doctoral dissertation, University of Siena, Chapter 2. Reproduced here without revision.

situation, there is no possibility of collusion. Therefore, the potential collision scenario is ruled out, and the pedestrian's lowest and highest outcomes become equivalent, $a = c$, resulting in:

$$c = a > e = g \tag{2}$$

For the AV, the highest payoff, f arises when the pedestrian yields and the AV proceeds. The lowest payoff, b corresponds to a potential collision. When the AV chooses to stop and yield, it receives the same payoff whether the pedestrian goes or stops, $d = h$. This leads to the following preference structure for DAV and AVI:

$$f > d = h > b \tag{3}$$

In contrast, AVII prioritizes the safety of the vehicle's occupants and is therefore more assertive in its behavior. While f remains the highest payoff, stopping and yielding ($d = h$) becomes least desirable. Because AVII deprioritizes external risk in favor of protecting passengers, the inconvenience or delay caused by yielding is weighted as more costly than the risk of a non-severe collision (b). This creates a distinct preference structure:

$$f > b > d = h \tag{4}$$

4.3 The No-Chicken Game: Pedestrian Strategy and DAV Predictability

We consider a simultaneous-move game of complete information, in which both the pedestrian and the DAV know each other's preferences. This interaction is modeled as a static Nash game, illustrated in Table 3.

Table 3. "No Chicken" Game Between Pedestrian and DAV

		DAV	
		Go	**Stop**
Pedestrian	**Go**	c, b	c, d
	Stop	e, f	g, h

Note. See Yoo [38], Chapter 2. Reproduced without revision.

The pedestrian's dominant strategy is to **Go**, exploiting the DAV's predictable behavior of always choosing **Stop**. Because the DAV is programmed to stop in any situation, its payoffs follow the preference ordering $f > d = h > b$. The pedestrian, aware of this deterministic behavior, has no incentive to yield, leading to a payoff structure where $a = c$ and $c > e = g$ in Equation 2. Thus, **Go** becomes the pedestrian's dominant strategy. Unlike a human pedestrian, the DAV cannot "win" the game of chicken due to its hard-coded commitment

to safety and legal compliance. Faced with the pedestrian's dominant move to Go, the DAV chooses to Stop, resulting in a stable outcome **(Go, Stop)** - the game's Nash equilibrium[9]. This outcome exemplifies the broader concern raised by Millard-Ball [28], identified as pedestrian supremacy: pedestrians exploit AV's predictability, gaining strategic power in mixed traffic environments. The predictable and risk-averse behavior of DAV eliminates strategic uncertainty, allowing pedestrians to exploit its safety programming.

4.4 A Strategic Game with Type I and Type II AVs

We now formally analyze the strategic interactions between a pedestrian and each of the two AV types - AVI (Type I) and AVII (Type II) - in the context of a simultaneous-move game with complete information. All players (the pedestrian, AVI, and AVII) are assumed to know their own and each other's preferences. While the pedestrian's preferences remain consistent across both AV types, the strategic dynamics vary based on the AV's risk attitude and ethical priorities. Only the AV's risk attitude for the pedestrian differs across the two types, and this distinction is reflected in the payoff matrix used in the analysis.

Pedestrian and AVI (Outsider Protection Priority) The pedestrian is risk-neutral towards AVI. Her payoffs satisfy $c > e = g > a$, where c is the highest payoff (AVI yields and pedestrian goes) and a is the lowest payoff (potential collision) in Equation 1.

AVI, in turn, is risk-averse towards the pedestrian. Its payoffs follow the structure $f > d = h > b$ in Equation 3. Although AVI is programmed to protect outsiders such as pedestrians, its ethical prioritization is not a guarantee of absolute safety, especially when other vulnerable road users like cyclists and motorcyclists are involved. Therefore, a residual risk of collision remains when pedestrians violate traffic rules.

The strategic interaction yields two pure-strategy Nash equilibria:

(1) **(Go, Stop)**: the pedestrian goes while AVI yields
(2) **(Stop, Go)**: the pedestrian yields while AVI proceeds

This scenario reflects a desirable outcome in mixed traffic, offering flexibility and reciprocity in movement. Importantly, the introduction of AVI mitigates the moral hazard issues associated with DAV and improves transportation efficiency.

Pedestrian and AVII (Insider Protection Priority) In contrast, the pedestrian is risk-averse toward AVII, anticipating a lower likelihood of being yielded to. However, her preferences remain the same as shown in Equation 1.

AVII is risk-neutral toward pedestrians and prioritizes passenger safety, favoring occupant protection in ethical dilemmas. Stopping is its worst outcome,

[9] The notion of Nash equilibrium for a strategic game models a steady state in which each player's beliefs about the other players' actions are correct, and each player acts optimally, given her beliefs.

wth payoffs structued as in Equatioin 4, making **Go** its dominant strategy. Given AVII's dominant strategy, the pedestrian chooses **Stop** to avoid harm, resulting in the Nash equilibrium **(Stop, Go)**. This reflects self-protective AV supremacy, where AVII deters pedestrian activity. While efficient for AVII, this dynamic undermines system efficiency by marginalizing human road users.

4.5 A Static Bayesian Game with AVI and AVII

Conitzer et al. [30] suggest that game-theoretic models with imperfect or incomplete information can extend moral solution concepts and better capture ethical concerns in multi-agent settings. In games of imperfect information, a player may not know the complete history of prior moves when making a decision. Since our model involves simultaneous move - where players choose their actions once and for all without observing others - we characterize our setting as one of incomplete information rather than imperfect information. To model this uncertainty formally, we introduce a static Bayesian game - the normal-form representation of a simultaneous-move game of incomplete information. In such a game, at least one player is uncertain about another player's type or payoff function.

In this section, we define the static Bayesian game structure and Bayesian Nash equilibrium as follows:

Definition 2 (Bayesian Game with Incomplete Information). *A Bayesian game consists of:*

- *A set of players*
- *A set of possible states (or types)*

For the pedestrian specifically*:*

- *A set of actions*
- *A belief distribution over the AV types (states)*
- *A Bernoulli payoff function that assigns expected utilities*

Assumptions The players are pedestrians and two distinct types of AVs: AVI (outsider protection priority) and AVII (insider protection priority). Each player chooses between two possible actions: Go and Stop. All players are regarded as rational decision-makers who choose the best action to maximize their expected utility, given their available information.

The AVs are fully informed about their own type and preferences. The pedestrian possesses common knowledge about the existence of two AV types - Type I (AVI), which prioritizes the protection of outsiders, and Type II (AVII), which prioritizes the protection of insiders—and understands each type's behavioral characteristics and payoff structure. However, the pedestrian cannot observe which specific type she is facing in any given encounter. This creates **type uncertainty**: while the pedestrian knows that both AV types exist and understands how each type behaves, she must form beliefs about whether the approaching vehicle is AVI or AVII in each interaction.

Bayesian Nash Equilibrium Under AV Type Uncertainty To model this type uncertainty formally, we characterize the interaction as a static Bayesian game. The pedestrian assigns a belief distribution over AV types: let $p \in [0, 1]$ denote the probability that the approaching AV is Type I, with probability $(1-p)$ assigned to Type II. Table 4 presents the payoff matrices corresponding to each AV type—the left panel shows payoffs when the AV is Type I, and the right panel shows payoffs when the AV is Type II.

To act rationally under uncertainty, the pedestrian computes her expected payoff (Bernoulli payoff) for each action by integrating across both possible AV types weighted by her beliefs. Formally, this expected payoff is computed as follows. The pedestrian's preferences are assumed to be consistent across interactions with both AV types, satisfying the condition in Equation 1. However, the AVs differ in their ethical priorities, reflected in distinct payoff structures. The outsider-protective AV (AVI) follows the preference ordering $f > d = h > b$, while the insider-protective AV (AVII) follows $f > b > d' = h'$, where both a and b are negative values representing potential collision costs. Because the payoffs are cardinal, these differences capture the ethical distinctions between AV types - specifically, whom they prioritize in conflict situations. For instance, both AV types face the same collision cost $b = -0.1$, but AVI receives a higher payoff $(d = 0)$ for yielding than AVII does $(d' = -0.2)$. This difference illustrates AVI's stronger commitment to outsider protection: AVI is willing to yield at no cost, while AVII incurs a penalty for doing so.

Table 4. Bayesian Payoff Matrix: Pedestrian vs. AV of Uncertain Type

Prob. p			
		AVI	
		Go	**Stop**
Pedestrian	**Go**	a, b	c, d
	Stop	e, f	g, h

Prob. $1-p$			
		AVII	
		Go	**Stop**
Pedestrian	**Go**	a, b	c, d'
	Stop	e, f	g, h'

Note. See Yoo [38], Chapter 2. Reproduced without revision.

Table 5 presents the pedestrian's expected payoffs under type uncertainty. To illustrate the computation, consider the case where the AV's action pair is (Go, Stop). If the pedestrian chooses Go, she receives payoff a with probability p (when facing AVI) and payoff c with probability $1 - p$ (when facing AVII). Her expected payoff is therefore: $p \cdot a + (1 - p) \cdot c = c - p(c - a)$. If she instead chooses Stop, her expected payoff is $p \cdot e + (1 - p) \cdot g = g + p(e - g) = g$, where the simplification follows from our assumption of $e = g$. Applying this procedure to all action combinations yields the complete expected payoff matrix shown in Table 5. Each column in Table 5 corresponds to an action pair chosen by the AV types. For example, the column labeled (Stop, Go) refers to a situation where AVI chooses Stop and AVII chooses Go.

Table 5. Expected Pedestrian Payoffs for AV Action Pairs

	(Go, Go)	**(Go, Stop)**	**(Stop, Go)**	**(Stop, Stop)**
Go	a	$c - p(c - a)$	$a + p(c - a)$	c
Stop	e	g	e	g

Note. See Yoo [38], Chapter 2. Reproduced without revision.

We define a Bayesian Nash Equilibrium in this context as a triple (a_p, a_1, a_2), where a_p is the pedestrian's action, a_1 is the action of AVI, and a_2 is the action of AVII such that the pedestrian's choice a_p is optimal given her beliefs and the actions (a_1, a_2), and each AV type's action is optimal given the pedestrian's choice a_p.

To simplify the analysis, we treat the two AV types as separate players in a three-player strategic game: the pedestrian's payoffs depend on both AVI's and AVII's actions (via Table 5). Each AV's payoff depends only on the pedestrian's choice (via the corresponding matrix in Table 4). The AVs do not interact or influence each other's payoffs.

In a Bayesian Nash Equilibrium, the pedestrian's action is a best response - based on her expected payoffs in Table 5 - to the pair of actions chosen by the two AV types. Simultaneously, each AV type's action is a best response to the pedestrian's choice, as specified by their respective payoff matrices in Table 4: the AVI responds according to the left panel, and the AVII to the right.

We now analyze whether specific action profiles satisfy these mutual best response conditions and therefore constitute Bayesian Nash Equilibria.
If the AVs choose (Go, Go), then the pedestrian's best response is Stop, since $e > a$ in Table 5. Given the pedestrian's choice of Stop, both AVI and AVII optimally choose Go, as shown in Table 4. Therefore, (Stop, (Go, Go)) is a Bayesian Nash equilibrium.
If the AVs choose (Stop, Stop), then the pedestrian's best response is Go, since $c > g$ in Table 5. However, given the pedestrian's action, AVI prefers Stop ($d > b$), while AVII prefers Go (since $b > d'$), as shown in Table 4. Thus, (Go, (Stop, Stop)) is not a Bayesian Nash equilibrium.
We check whether (Go, (Stop, Go)) could be an equilibrium. In this case, given the AVs' actions (Stop, Go), the pedestrian receives a from AVI and c from AVII, as shown in Table 4. The pedestrian's expected payoff from choosing Go is $a + p(c - a)$, as shown in Table 5. This is better than Stop if and only if $a + p(c - a) \geq e$. This is,

$$p \geq \frac{e - a}{c - a} \tag{5}$$

Therefore, (Go, (Stop, Go)) is a Bayesian Nash equilibrium if and only if this condition is satisfied.

Other configurations can be ruled out similarly: (Stop, (Go, Stop)) and (Stop, (Stop, Go)) are not equilibria. Because when the pedestrian chooses Stop,

both AV types strictly prefer Go in response, as shown in Table 4, violating the best-response condition.
(Go, (Go, Stop)) is not a Bayesian Nash equilibrium because, when the pedestrian chooses Go, AVI prefers Stop over Go (since $d > b$) in Table 4, violating its best-response condition.
(Go, (Go, Go)) has already been ruled out because it leads the pedestrian to choose Stop, which does not support Go as a best response for her.
(Stop, (Stop, Stop)) also fails because the pedestrian prefers Go as a best response when both AV types choose Stop.

5 Results

The analysis of Bayesian Nash Equilibria reveals that the proportion of AVI on the road determines the pedestrian's optimal strategy and, consequently, the strategic balance of the mixed traffic environment.

If the proportion of AVI is less than a threshold value $\bar{p}$, then the pedestrian's best response is always to Stop - yielding to both AVI and AVII. In this case, a unique Bayesian Nash equilibrium exists:

$$(\text{Pedestrian}, \text{AVI}, \text{AVII}) = (\text{Stop}, \text{Go}, \text{Go})$$

This outcome reflects AV supremacy, where both AV types dominate pedestrian behavior.

If the proportion of AVI is greater than or equal to a threshold value $\bar{p}$, then there are two Bayesian Nash equilibria:

$$(\text{Pedestrian}, \text{AVI}, \text{AVII}) = (\text{Stop}, \text{Go}, \text{Go}) \quad \text{and} \quad (\text{Go}, \text{Stop}, \text{Go})$$

These equilibria are both stable and reflect coexisting social norms.[10] In practice, which equilibrium becomes focal may depend on local expectations, infrastructure, and cultural norms. The equilibrium threshold $\bar{p}$ is given by

$$\bar{p} = \frac{e - a}{c - a} \tag{6}$$

Here, a is the pedestrian's payoff in a potential collision (assumed negative), c is the payoff when the AV yields, and e is the payoff when the pedestrian yields. Given that $c > e > a$ and $a < 0$, the threshold satisfies $0 < \bar{p} < 1$.

As the cost of potential collision (a) decreases in magnitude (i.e., the collision becomes less severe or less costly), $\bar{p}$ decreases. This means fewer AVI are needed to maintain safe pedestrian behavior. Conversely, as the cost of collision rises (i.e., a becomes more negative), more AVI are required to deter risky pedestrian behavior and avoid AVII dominance.

[10] In games with multiple Nash equilibria, certain outcomes may be more psychologically or culturally salient, and therefore more likely to attract players' expectations. These preferable outcomes are referred to as focal equilibria, and often function as social norms in real-world strategic interactions [40].

As the benefit of AV yielding increases (i.e., a higher c), pedestrians derive more satisfaction from being yielded to. This reduces the marginal value of adding more AVI to the road, because the existing AVI already makes AV behavior feel safe and rewarding. As a result, fewer additional AVI are needed, and the equilibrium threshold $\bar{p}$ decreases. Intuitively, the more rewarding the pedestrian's interaction with existing AVI, the less pressure to increase the proportion of AVI, because pedestrian trust is already high and pedestrian supremacy becomes a risk if AVI dominates the traffic. When the benefit of AV yielding (c) decreases, a higher proportion of AVI is needed. As pedestrians become less satisfied or feel less rewarded by AVI yielding behavior, society requires more AVI to restore trust and protect pedestrian safety. In this context, increasing the number of AVI helps prevent AVII supremacy, where insider (or self-protective) AVs dominate the traffic environment.

As the pedestrian's payoff from yielding (e) decreases, the equilibrium threshold $\bar{p}$ also decreases. In this case, stopping becomes less rewarding, making pedestrians more likely to exploit the presence of AVI by choosing to cross. To prevent this strategic exploitation and the resulting risk of pedestrian supremacy, fewer AVI are needed to ensure pedestrians opt to stop. Fewer AVI implies a higher proportion of AVII on the road, and AVII's assertive behavior serves as a deterrent, promoting safer pedestrian behavior in mixed traffic. Conversely, as e increases, stopping becomes more attractive to pedestrians, especially in environments where AVII are prevalent. In such cases, more AVI are required to maintain balance and prevent the rise of self-protective AV (AVII) supremacy, where assertive AV behavior disproportionately suppresses pedestrian agency.

It is noteworthy that AVII consistently chooses to Go in all equilibrium scenarios.[11] This reflects its programmed prioritization of passenger safety and aligns with current trends in AV design. Meanwhile, introducing AVI serves to mitigate the ethical concerns associated with exclusive insider protection and provides a balancing mechanism between AV supremacy and pedestrian supremacy in a mixed traffic environment.

6 Conclusion

This study proposes a new ethical reasoning framework for AV decision-making by distinguishing between insider and outsider protection priorities. We introduce and formalize two distinct AV types: AVI (outsider-protection) and AVII (insider-protection). These AVs embody heterogeneous moral preferences and offer a scalable design strategy that accommodates the diverse interests of traffic participants.

Our static Bayesian game model demonstrates that the mere existence of AVI is not sufficient. Instead, maintaining a critical proportion of AVI on the road

[11] While AVII's dominant strategy of *Go* may seem counterintuitive, this analysis specifically models ethical dilemma situations in which a collision is unavoidable. Under normal traffic conditions, AVII would be expected to obey legal norms and yield at marked crosswalks and traffic lights.

significantly improves transportation efficiency and pedestrian safety in mixed traffic systems. Importantly, this design helps discourage moral hazard behavior by pedestrians, particularly in environments dominated by fully risk-averse AVs, such as DAV.

From a policy perspective, the key challenge lies in determining the optimal proportion $\bar{p}$ of AVI that balances pedestrian safety with overall transportation efficiency. Policymakers may aim to keep p close to, but below 1, to avoid pedestrian supremacy while still preserving the deterrent value of AVII's assertiveness.

Furthermore, strategies that reduce $\bar{p}$, such as lowering pedestrian collision costs (a) through urban design or increasing the reward from AV yielding (c), can reduce the number of AVI needed to maintain equilibrium. This approach offers a cost-effective path to Pareto-efficient outcomes in ethically complex and behaviorally diverse traffic environments.

Acknowledgments. The authors thank Nicola Dimitri for introducing the Bayesian game concept, and Rune Nyrup for valuable suggestions and insightful comments on the payoff structures in the game model.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Toulouse School of Economics Homepage, *Whose Life Should Your Car Save?* https://www.tse-fr.eu/article/whose-life-should-your-car-save , last accessed 2021/04/20
2. OECD Homepage, https://doi.org/10.1787/eedfee77-en , last accessed 2023/01/08
3. Mecacci, G., Calvert, S.C., Santoni de Sio, F.: Human–machine coordination in mixed traffic as a problem of Meaningful Human Control. AI & Society **38**, 1151–1166 (2023). https://doi.org/10.1007/s00146-022-01605-w
4. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science **352**(6293), 1573–1576 (2016). https://www.science.org/doi/abs/10.1126/science.aaf2654
5. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. Proceedings of the IEEE **107**(3), 502–504 (2019). https://doi.org/10.1109/JPROC.2019.2897447
6. Greene, J.D.: Our driverless dilemma. Science **352**(6293), 1514–1515 (2016). https://www.science.org/doi/abs/10.1126/science.aaf9534
7. Millar, J.: You should have a say in your robot car's code of ethics. Wired (2014). https://www.wired.com/ , last accessed 2022/09/01
8. Millar, J.: An ethical dilemma: when robot cars must kill, who should pick the victim? Robohub (2014). https://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/, last accessed 2022/09/01
9. Lin, P.: The ethics of autonomous cars. The Atlantic (2013). https://www.theatlantic.com/technology , last accessed 2022/09/01

10. Floridi, L., Cowls, J., Beltrametti, M., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds & Machines **28**, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
11. March, C.: Strategic interactions between humans and artificial intelligence: lessons from experiments with computer players. Journal of Economic Psychology **87**, 102426 (2021). https://doi.org/10.1016/j.joep.2021.102426
12. Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3306618.3314289
13. Coeckelberg, M.: AI Ethics. The MIT Press, Cambridge (2019)
14. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, New York (2009)
15. Lin, P., Jenkins, R., Abney, K.: Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence. Oxford University Press, New York (2017)
16. Wallach, W., Franklin, S., Allen, C.: A conceptual and computational model of moral decision making in human and artificial agents. Topics in Cognitive Science **2**(3), 454–485 (2010). https://doi.org/10.1111/j.1756-8765.2010.01095.x
17. Vamplew, P., Dazeley, R., Foale, C., et al.: Human-aligned artificial intelligence is a multiobjective problem. Ethics and Information Technology **20**, 27–40 (2018). https://doi.org/10.1007/s10676-017-9440-6
18. Nallur, V.: Landscape of machine-implemented ethics. Science and Engineering Ethics **26**, 2381–2399 (2020). https://doi.org/10.1007/s11948-020-00236-y
19. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in machine ethics: a survey. ACM Computing Surveys **53**(6) (2021). https://doi.org/10.1145/3419633
20. Bonnemains, V., Saurel, C., Tessier, C.: Embedded ethics: some technical and ethical challenges. Ethics and Information Technology **20**, 41–58 (2018). https://doi.org/10.1007/s10676-018-9444-x
21. Kirkpatrick, K.: The moral challenges of driverless cars. Communications of the ACM **58**(8) (2015). https://doi.org/10.1145/2788477
22. Luetge, C.: The German ethics code for automated and connected driving. Philosophy & Technology **30**, 547–558 (2017). https://doi.org/10.1007/s13347-017-0284-0
23. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J., Rahwan, I.: The Moral Machine experiment. Nature **563**, 59–64 (2018). https://doi.org/10.1038/s41586-018-0637-6
24. Awad, E., Dsouza, S., Shariff, A., Rahwan, I., Bonnefon, J.-F.: Universals and variations in moral decisions made in 42 countries by 70,000 participants. Proceedings of the National Academy of Sciences **117**(5), 2332–2337 (2020). https://doi.org/10.1073/pnas.1911517117
25. Lin, P., Abney, K., Bekey, G.A.: Robot Ethics: The Ethical and Social Implications of Robotics. The MIT Press, Cambridge (2012)
26. Etzioni, A., Etzioni, O.: Incorporating ethics into artificial intelligence. The Journal of Ethics **21**, 403–418 (2017). https://doi.org/10.1007/s10892-017-9252-2
27. Zhu, Y., et al.: A moral decision-making study of autonomous vehicles: expertise predicts a preference for algorithms in dilemmas. Personality and Individual Differences **186**, 111356 (2022). https://doi.org/10.1016/j.paid.2021.111356
28. Millard-Ball, A.: Pedestrians, autonomous vehicles, and cities. Journal of Planning Education and Research **38**(1), 6–12 (2018). https://doi.org/10.1177/0739456X16675674

29. Braithwaite, R.: *Theory of Games as a Tool for the Moral Philosopher*. An inaugural lecture delivered in Cambridge on 2 December 1954. Cambridge University Press, Cambridge (1955)
30. Conitzer, V., Sinnott-Armstrong, W., Schaich Borg, J., Deng, Y., Kramer, M.: Moral decision making frameworks for artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence **31**(1) (2017). https://doi.org/10.1609/aaai.v31i1.11140
31. Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V.R., Yang, Q.: Building ethics into artificial intelligence. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18) (2018). https://arxiv.org/abs/1812.02953
32. de Melo, C.M., Marsella, S., Gratch, J.: Human cooperation when acting through autonomous machines. Proceedings of the National Academy of Sciences **116**(9), 3482–3487 (2019). https://doi.org/10.1073/pnas.1817656116
33. Di, X., Chen, X., Talley, E.: Liability design for autonomous vehicles and human-driven vehicles: a hierarchical game-theoretic approach. Transportation Research Part C: Emerging Technologies **118**, 102710 (2020). https://doi.org/10.1016/j.trc.2020.102710
34. Nyholm, S., Smids, J.: Automated cars meet human drivers: responsible human-robot coordination and the ethics of mixed traffic. Ethics and Information Technology **22**, 335–344 (2020). https://doi.org/10.1007/s10676-018-9445-9
35. Horty, J.F.: Skepticism and floating conclusions. Artificial Intelligence **135**(1), 55–72 (2002). https://doi.org/10.1016/S0004-3702(01)00160-6
36. Bonnefon, J.-F.: Reinstatement, floating conclusions, and the credulity of mental model reasoning. Cognitive Science **28**(4), 621–631 (2004). https://doi.org/10.1207/s15516709cog2804_6
37. Yoo, D.H.: A logical approach in autonomous vehicle ethics: the skeptical reasoning in dilemma. AI Ethics **5**, 3069–3078 (2025). https://doi.org/10.1007/s43681-024-00616-1
38. Yoo, D.: The ethics of artificial intelligence from an economics perspective: logical, theoretical, and legal discussions in autonomous vehicle dilemma. Ph.D. thesis, University of Siena, Italy (2023). https://doi.org/10.25434/yoo-dae-hyun_phd2023
39. Bowles, S., Halliday, S.: Microeconomics: Competition, Conflict and Coordination. Oxford University Press, Oxford (2021)
40. Schelling, T.: The Strategy of Conflict. Harvard University Press, Cambridge (1960)