# Human Preferences in Moral Decision Making for Autonomous Vehicles

Bishal Thapa
*Department of Computer Science*
Texas State University
b_t220@txstate.edu

Henry Griffith
*Dept. of Engineering*
*San Antonio College*
hgriffith5@alamo.edu

Heena Rathore
*Department of Computer Science*
*Texas State University*
heena.rathore@txstate.edu

***Abstract*—There has been tremendous growth in Autonomous Vehicles (AVs) recently, yet the moral decision-making capabilities remain a crucial challenge that needs to be addressed. Addressing this issue is pivotal for gaining societal trust towards AVs, as the decision-making of the AVs will impact human life in a significant manner. Incorporating human preferences in the decision-making of AVs ensures ethical decisions along with societal acceptance of decisions made under moral uncertainty. In this paper, we propose integrating human preferences into Reinforcement Learning (RL) to guide AVs to make human-like decisions. We use the Bradley-Terry (BT) model to incorporate human preferences and perform pairwise comparisons on the moral machine framework of AVs. This approach of considering human preference adds a layer of explainability to the decisions and enhances the significance of the results for real-world applicability. The results show the decision-making capability of RL agents could be improved by embedding human preferences and the decisions made by AVs align closely with those of humans. These results suggest that while the decision on AVs is likely to be controversial, the incorporation of human preferences can foster societal trust in decisions made by AVs.**

***Index Terms*—Autonomous Vehicles, Reinforcement Learning, Bradley-Terry, Human Preferences**

## I. Introduction

Autonomous Vehicles (AVs) are increasingly making their mark in the automotive industry, owing to the promise of effectiveness in navigation and the potential to significantly reduce road accidents [1]. However, a major hurdle to these advancements is a critical challenge that tests the capability of AVs to make ethical decisions in cases of moral uncertainty [2]. If AVs are better adept at making ethical moral decisions, it would substantially elevate public trust and confidence, which would further foster the development of AVs.

Recently, reinforcement learning (RL) based models have been designed to address the capability of agents to make decisions in scenarios of moral uncertainty [3], [4]. These works predominantly focus on utilitarianism and deontology as two principal ethical frameworks to guide the decision-making capability of AVs. While it's commendable to consider ethical perspectives, the challenge arises from the inherent conflicts and subjective interpretations of these theories. One issue is the assignment of numerical scores (in the form of credence values) to actions based on moral theories like deontology, and utilitarianism [5]. Each theory carries its own set of principles and priorities, leading to potential contradictions when trying to apply them uniformly. For instance, deontology emphasizes rules and obligations, while utilitarianism focuses on maximizing overall happiness or well-being. These two perspectives can conflict when determining the ethical course of action in each scenario. Furthermore, the assignment of specific numerical values based on credence values to actions may oversimplify complex ethical dilemmas.

The challenge of validating the effectiveness and generalizability of ethical decision-making algorithms across diverse scenarios is a critical concern in the development of AVs. By considering human preferences, we can simulate more realistic scenarios that reflect how human drivers perceive and interact with situations under morally uncertain situations [6]. In this paper, we incorporate human preferences using the Moral Machine framework, specifically designed to gather such preferences [8]). No prior work has utilized these vast human preferences directly in AVs for ethical decision-making. We utilize the Bradley-Terry model within the Moral Machine framework [7] to ascertain credence values. These values, infused with human preferences, are then integrated into the RL model outlined by [5]. This integration enhances the RL agent's decision-making process, making its choices more pertinent and aligned with human values.

## II. Related Work

The paper [14] delves into the importance of alignment of AI with human values. It explores the interplay of the normative and the technical aspects of this alignment along with its inherent complexity. The author discusses different approaches to AI alignment, by stating the importance of clear alignment goals and argues whether AI should align with human instructions, intentions, preferences, desires, interests, or values. For this purpose, the author explores the merits and demerits of these approaches by considering different research and exploring philosophical principles like utilitarianism and Kantian. The paper discusses how the central challenge is to come up with principles for true alignment of AI that are considered fair rather than finding the correct moral principle. This paper serves as an important reference to understand the integration of different principles and values into AI.

The authors in [11] discuss the integration of different ethical frameworks for AV control algorithms. They incorporated philosophical knowledge from deontology, and con-

sequentialism as guiding principles for AVs to meet societal expectations and ethical norms i.e., accident avoidance and adherence to traffic laws. This paper demonstrates how embedding philosophical principles as system constraints and objectives can drive vehicle control to align with societal expectations. The authors implement deontological principles for the development of constraints in the vehicular system, consequentialism for the construction of the objective function, and implemented Model Predictive Control (MPC) to incorporate these ethical theories into a control algorithm. MPC aims to minimize the costs guided by consequentialism and constraints guided by deontology to determine the optimal path for the vehicle. The experimental results demonstrate that varying weights on costs or constraints represented by different ethical theories significantly affect the behavior of the test vehicle. They conclude the necessity of integrating different ethical principles for responsible programming of AVs. However, the study doesn't address the inherent difficulty of balancing between these subjective ethical frameworks.

Authors in [10] deal with understanding the human preferences on the ethical dilemma (ED) of who the AVs should protect in cases where harm is unavoidable. The paper provides a unique view of potential AV consumers and their standpoint on risk perception regarding AVs. The authors also offer different technical, legal, and ethical challenges from the perspective of potential AV adopters. Two important experimental studies were performed with broad consumer samples which depicted that people associated EDs with the highest risk. These empirical studies also revealed that consumers considered EDs to be the most important issue facing EVs compared to various legal and technical issues. Overall, the findings from the paper highlight the gravity of the ethical dilemma for broader acceptance of AVs. While the study explores different aspects of consumer perceptions towards AVs, the results are based on a specific consumer sample that does not represent a diverse demographic spectrum and could limit the generalizability of the results.

Andrea et. al [6] investigated an approach to build AI agents capable of making human-like decisions, especially in situations requiring difficult tradeoffs or moral uncertainty. The authors explored the Multi-Alternative Decision Field Theory (MDFT) as a basis for developing an AI orchestrator called MDFT-Orchestrator (MDFT-O). This orchestrator tries to make decisions in a constrained environment modeled using deontological and consequentialist frameworks in a way that balances between reaching the goal and satisfying the ethical constraint. They used Markov Decision Processes (MDPs) for modeling the decision environment. MDFT-O consistently performed better than other models like Greedy Orchestrator and Weighted Average Orchestrator. They further validated their model using experimental human preference data collected through Amazon Mechanical Turk. Despite the impressive performance of the orchestrator, the small sample size of 185 participants on Amazon Mechanical Turk limits the effectiveness of the validation.

Bogosian [12] proposes a computation framework for an AI system to address the problem of moral disagreement among different moral philosophies. The author suggests a framework where machines are morally uncertain and thus programmed to make decisions by considering multiple moral theories and their respective priorities according to their plausibility. The author takes reference from MacAskill's metanormative theory approach proposed in [13]. MacAskill explained the issue of moral uncertainty as a voting mechanism that involves computing the expected choice-worthiness of action by considering credence towards different moral theories. Bogosian takes a similar approach and applies MacAskill's philosophical concept to implement moral uncertainty in AI systems.

## III. Methods

### A. Background on RL Model Under Moral Uncertainty

Authors in [5] introduced a RL based approach to tackle the moral uncertainties by addressing the issue of comparability of different ethical theories, as explored by MacAskill [13]. They particularly focused on two normative theories: utilitarianism and deontology. Utilitarianism is a consequentialist ethical theory that solely focuses on the consequences of actions. Deontology, on the other hand, emphasizes adhering to moral rules to distinguish right from wrong. Each theory is represented by a certain degree of belief towards that theory defined as credence. The authors implemented a voting mechanism as a means to integrate different ethical theories in decision-making by defining a choice-worthiness function $W_i$, analogous to a reward function for each ethical theory. The reward function for these two theories is represented as a credence-weighted sum of the choice-worthiness according to each theory, where $C_i$, denotes the credence level of each theory, $a$ represents the action state, $s$ represents the current state, and $s'$ represents the next state:

$$R(s, a, s') = \sum_i C_i W_i(s, a, s') \tag{1}$$

The RL agents must balance between multiple choice-worthiness functions instead of simply maximizing a single reward function. So, they defined $Q_i(s, a)$ to represent the discounted sum of the choice-worthiness function $W_i(s, a, s')$ for each theory $i$ by taking action $a$ at state $s$ under given policy $\pi$.

$$Q_i(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t W_i(s_t, a_t, s_{t+1}) \middle| s_0 = s, a_0 = a\right] \tag{2}$$

where $\gamma_i \in [0, 1]$ represents the discount factor. The problem of incomparable choice-worthiness functions across different theories is discussed in the paper through the principle of Proportional Say, where the influence of theory is adjusted proportionally to its credence and not through the choice-worthiness function. For this purpose, the authors devise two voting mechanisms: Nash voting and Variance voting.

- Nash voting tends to make decisions through tactical voting rather than reflecting the true preferences of the

theory. Each theory is associated with an RL agent, and it seeks to maximize the choice worthiness for that theory. Theories are provided with an equal voting budget and hence follow the principle of proportional say. The agents have to pay the cost for each vote they cast which is proportional to the size of their vote. These votes are then scaled proportionally with their credence values and theories use them to vote for or against the available actions. The voting budget can be exhausted after which the votes of that agent are ignored.

- Variance voting or Variance-Sarsa involves adjusting votes with a focus on the variability of the outcomes of different actions and does not suffer from stakes insensitivity. Variance voting mechanism overcomes the issues of Nash voting by variance-normalizing the preferences of the theories. It reflects the risk associated with the actions and makes decisions with risk-aware assessment i.e., actions with lower variance. The on-policy Q-values of each theory are learned using Sarsa and are considered as the preference of that theory. These preferences are then converted into votes through variance normalizing by the expected value of variance ($\sigma_i^2$) across timesteps using 4.

$$\mu_i(s) = \frac{1}{k} \sum_a Q_i(s, a), \tag{3}$$

where $k$ represents the actions over discrete action space. Then;

$$\sigma_i^2 = E_{s \sim S} \left[ \frac{1}{k} \sum_a (Q_i(s, a) - \mu_i(s))^2 \right] \tag{4}$$

Nash voting and variance voting mechanisms exhibit distinct impact on the ethical decision-making process. While Nash voting tends to produce more polarized decisions based on the dominant ethical theory, variance voting demonstrates a more cooperative approach by considering the variability of outcomes. Nash voting learns competitive strategies to learn the voting policies, whereas variance voting uses cooperative integration of the theories to learn value functions that are used for voting. This leads to more balanced decisions, particularly in scenarios where the stakes or consequences of actions vary significantly. Furthermore, both of these voting mechanisms exhibit distinct properties and limitations which can be discussed through Arrow's desirability axioms. It establishes essential properties for voting mechanisms, which include non-dictatorship, pareto efficiency, and independence of irrelevant alternatives (IIA). Non-dictatorship emphasizes that the outcomes should reflect the input from all voting theories, so that no outcome is dictated by a single theory. Pareto efficiency prefers that an action should not be chosen if it is not preferred across by any theory. And IIA asserts that adding a new irrelevant action should not influence the decision-making process. Both voting mechanisms discussed in the paper adhere to only some of these properties. Nash voting does not guarantee Pareto efficiency, while variance voting doesn't not consistently satisfy the IIA. As Pareto efficiency is more desirable, as discussed by MacAskill, variance voting may be considered more beneficial. The other limitation of Nash voting is it's stakes insensitivity, where increasing stakes of one theory doesn't influence the overall decision. Furthermore, Nash voting also exhibits no compromise flaw where if no single theory chooses an action as the most preferred, that action cannot be chosen even if it is the best compromise action.

### *B. RL Environment*

The RL state $(s)$ is characterized by several elements within a grid environment. These elements include the presence of a switch, the spatial distribution of people represented through one-hot encoding (represented as $X$), and additional information regarding one's beliefs in a particular theory using credence values, as well as the count of people situated on the track as shown in Fig. 1. The agent actions are defined as up, down, left, and right. The agent's decision-making process involves navigating this grid-based environment based on the current state, with the goal of optimizing its performance over time through the selection of appropriate actions.

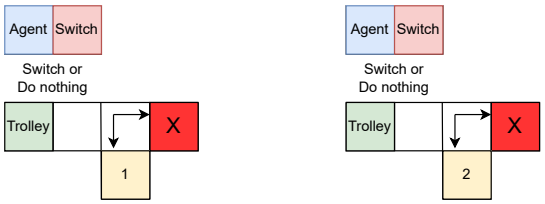



Fig. 1: Illustration of Trolley Problem Scenarios: (a) Classic Scenario and (b) Modified Double Scenario

In this work, we considered two scenarios presented in [5] as foundational scenarios to evaluate the performance of the RL agent with human preference as credence. The initial simulation involves a classic trolley dilemma as depicted in Fig. 1 (a), where a trolley is moving toward a group of $X$ individuals. The agent can flip the switch to divert the trolley to an alternate path and cause a collision with one individual instead, thereby saving X individuals from harm. The second simulation involves an adaptation of the double trolley problem, also presented in [5], where instead of one individual on the alternate path, there are two individuals on the track as depicted in Fig. 1 (b). We removed the action of pushing a large man to the path of the trolley, as this depicts situation which is highly improbable and unrealistic. We refined the double scenario for practical relevance by considering a more realistic "2 v X" scenario where decisions involve prioritizing saving two over X number of individuals. The severity of the actions from utilitarianism and deonotological perspectives, is quantified through weights assigned to different actions in these two scenarios. These values are hard-coded in the program and are presented in Table I and II.

| | Crash into 1 | Crash into X |
|---|---|---|
| **Utilitarianism** | -1 | -X |
| **Deontology** | -1 | 0 |

TABLE I: Preferences in the classic trolley problem.

| | Crash into 2 | Crash into X |
|---|---|---|
| **Utilitarianism** | -2 | -X |
| **Deontology** | -1 | 0 |

TABLE II: Preferences in the double trolley problem.

### C. Human Preferences and Credence Generation

*1) Collecting Human Preferences:* To embed human preferences into AVs, we employed the "Moral Machine Experiment" framework [8]. This platform collected a comprehensive dataset comprising 40 million decisions in ten languages from 233 countries. Participants were presented with morally ambiguous scenarios and asked to indicate their preferred choice. Specifically, we extract pairwise results for a classic scenario and double scenario from this framework to feed into the RL model. In the classic scenario, participants were asked to save either one individual or a group of $X$ individuals (where $X \in [2-5]$) analogous to the classic trolley as depicted in Fig. 1 (a). In the double scenario, participants are asked to save either two individuals or $X$ individuals (where $X \in [3-5]$ analogous to the modified double trolley as depicted in Fig 1 (b). Fig. 2 shows an example of a 1 v 3 scenario, where swerving is considered a utilitarian approach, and staying in the lane is considered a deontological approach.

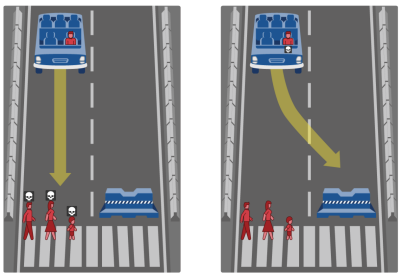


Fig. 2: Example Scenario from the Moral Machine [8]

*2) Bradley-Terry model for pairwise comparison:* The BT model is a widely used probability model for pairwise comparison analysis, providing a framework for deriving the score values or strengths of the choices in different scenarios of an experiment. The BT model assigns strength values to each item based on their comparisons. The model depends upon generating strength parameters based on observed comparisons [7]. This parameter can be generated using methods like maximum likelihood estimation. The probability model of pairwise comparison of outcomes $(i, j)$, where $\beta$ is a strength parameter of a particular choice, and $i$ is likely to be chosen over $j$ can be modeled as:

$$p_{ij} = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}} \tag{5}$$

We adapted this BT model to reflect the relative strengths of the decision for the classic and modified double scenario.

## IV. Results and Performance Analysis

### A. Calculation of Credence using BT model

Fig. 3 and Fig. 4 present a scenario analysis of moral decision-making from data gathered through the moral machine experiment. Fig. 3 depicts the total number of cases where participants were prompted to choose between saving one individual over $X$ individuals (where $X \in [2-5]$). The strength of the preference for saving $X$ individuals over one was calculated using 5, as described earlier in section III C. This strength is depicted by the red line in the chart and represents the degree to which human preference aligns with the utilitarianism principle. Fig. 4 extends the analysis by exploring scenarios in a double trolley problem, by addressing human preferences in 2 v X cases, where X varies from 1 to five. Similar to Fig. 3, it presents total cases of 2 v X from the moral machine experiment, normalizes the data by calculating the probability of saving X in each scenario, and generates the strength values of saving X using the BT model as depicted by the red line in the figure. The credence values is a critical aspect of our model that binds the human preferences with the model's decision. We generated the credence values for each scenario (e.g., 1v2, 1v3, etc) separately by filtering the scenarios from the "Moral Machine Experiment" that depicted decisions made in accordance with either Utilitarianism or Deontology. By deriving these values from empirical human preference data using the Bradley-Terry model, we aim to achieve greater transparency and improved alignment with human moral judgments which is a step towards making ethical decisions.

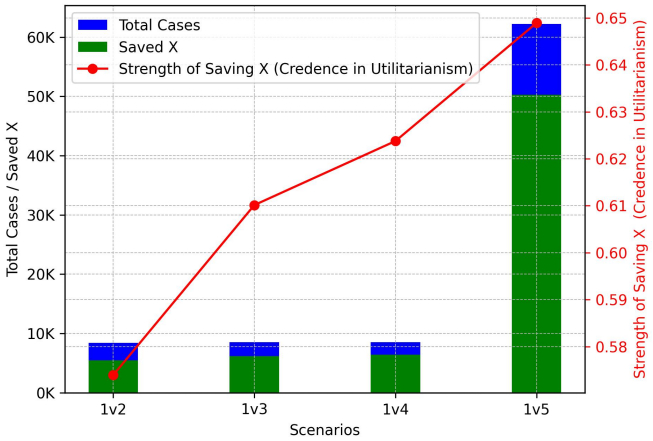



Fig. 3: Credence for Utilitarianism in 1 v X scenario

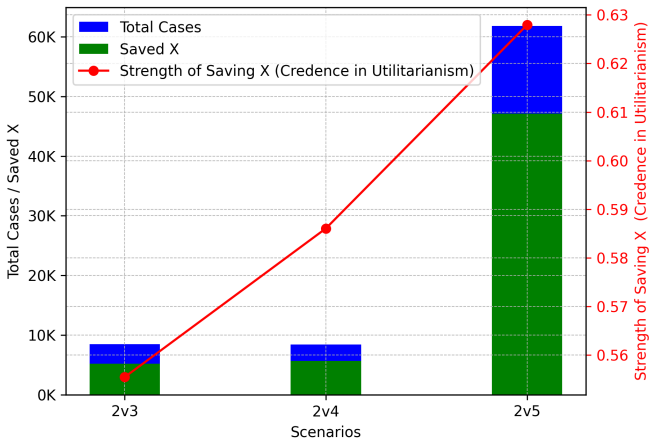



Fig. 4: Credence for Utilitarianism in 2 v X scenario

### B. Integrating Human Preferences in RL Framework

We simulated RL agents with the Nash and Variance voting [5] for classic and modified double scenarios. The experiments were performed on two distinct environmental setups: a sequential environment, and a non-sequential environment. The decision in a sequential environment is not isolated and requires the RL model to make multiple decisions (2 in our case) before the episode ends. The stakes of future decisions are unknown and hence the model needs to perform the strategic management of the voting budget under Nash voting. Conversely, the decisions in the non-sequential environment are isolated, and each episode requires the model to make a single decision. Fig. 5 and Fig. 6 represent the decision behavior of the agent for the classic scenario under the Nash voting and variance voting mechanisms in 1 v X scenarios.

*1) Classic Scenario:*

*a) Nash Voting:* Fig. 5(a) represents the decision-making behavior of the RL agent under random credence in a non-sequential environment. It shows that the agent tends to do nothing when the deontology credence value is above 50% regardless of the number of people on the track. Conversely, for cases where deontology credence is less than 50%, the agent tends to prefer the "switch" option indicating a preference in the utilitarian principle to minimize the overall harm. Fig. 5(b) explores the decision-making behavior of the agent where credence values are generated through human preference as discussed in Section III.C. The vertical lines mark the exact deontology credence value generated through human preferences for specific scenarios in the classic environment. The figure extends up to 50% deontology credence values for exploring the 1 v 1 situations where both ethical theories hold equal credence as a baseline.

Fig. 5(c) represents the decision-making behavior of the RL agent involved in sequential decision scenarios. It shows an interesting behavior of agents where the agents demonstrate stake insensitivity. Even with lower deontology credence, we can see situations where the agents tend to prefer doing nothing when we expected it to switch. But, regardless of the environment setup (sequential or non-sequential), the credence preferred by humans is still the same for a particular scenario. It can be seen that when the number of individuals on track exceeds 1, then the agent's decision under human preference credence is always to switch. This is consistent with the non-sequential decision-making depicted in Fig. 5(b). Therefore, in the classic environment under Nash voting, we found that the agent's decision under human preferred credence consistently favored switching the trolley depicting the dominance of the utilitarian principle. A distinct shift in decision-making can be observed when the decisions are compared to the original study, where the agent preferred to switch in approximately half of the scenarios. However, it was found that when agents were provided with credence values generated by considering human preferences, they consistently preferred to switch, reflecting human behavioral tendencies in scenarios analogous to the classic trolley dilemma.

*b) Variance Voting:* Figs. 6(a) and (b) represent the behavior of the model for the classic scenario under the variance voting mechanism, with the agent's decision influenced by credence generated randomly and human preferred credence respectively. As evident from Fig. 6, the decisions of RL agents following variance voting are not as one-sided decisions as Nash voting. This is because variance voting involves adjusting votes with a focus on the variability of the outcomes of different actions and does not suffer from stakes insensitivity as in Nash voting. Variance voting tries to find balance in impact by the varying outcomes, which helps in generating consistent decision-making regardless of the sequential or non-sequential scenario. As illustrated in Fig. 6(a), the agent consistently chooses to do nothing once the deontology credence exceeds approximately 60%. In contrast, agents with human-preferred credence consistently choose to switch for all cases except when the number of people on tracks is two, as depicted in Fig. 6(b).

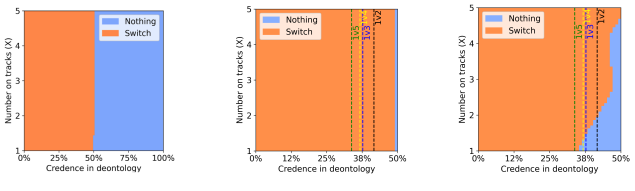
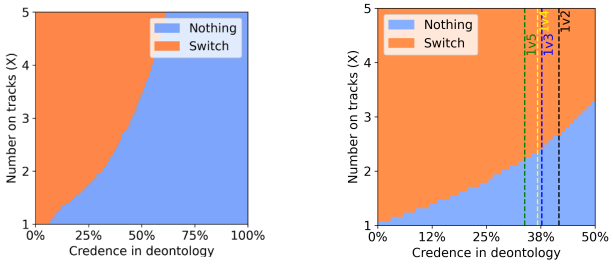
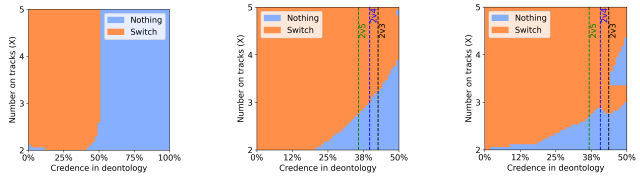



Fig. 5: Classic Environment Nash Voting (a) Random credence, (b) Human preference credence in Non-Sequential environment, (c) Human preference credence in Sequential environment



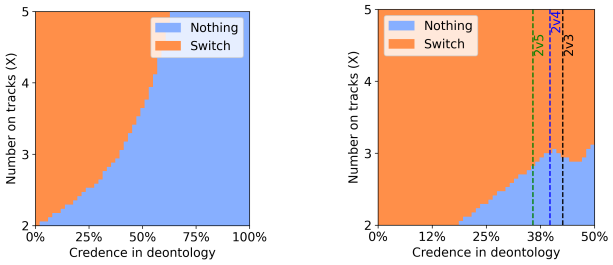

(a) (b)

Fig. 6: Classic Environment Variance Voting (a) Random credence, (b) Human preference credence




(a) (b) (c)

Fig. 7: Modified Double Environment Nash Voting (a) Random credence (b) Human preference credence in Non-Sequential environment, (c) Human preference credence in a Sequential environment




(a) (b)

Fig. 8: Modified Double Environment Variance Voting (a) Random credence (b) Human preference credence

*2) Modified Double Scenario:*

*a) Nash Voting:* Fig. 7(a) represents the decision behavior of the agent for the modified double scenario under the Nash voting mechanism with random credence. It shows similar results to the classic environment, with the agent preferring to do nothing when the deontology credence value is above 50%. However, the decision boundary is not as clearly defined as in the classic environment and the decision is leaning slightly more towards doing nothing. Fig. 7(b) and Fig. 7(c) depict the decisions of an agent under credence derived using our method in a non-sequential and sequential environment respectively. When the number of individuals on the track is two, the decision of the agent is more inclined towards doing nothing as the deontological credence surpasses approximately 20%, even though the scenario presents equal individuals on either track. This can be explained by the weight assigned by the ethical principles to different actions as depicted in Table I and Table II. The deontological principle focuses on the morality of action as right or wrong regardless of the number of people on the track. So, both actions of "Crash into 1" and "Crash into 2" receive -1 weight under deontology highlighting that deontology doesn't care about the outcome but the action. Utilitarianism, on the other hand, evaluates morality through outcomes. So, under this principle, "Crash into 1" and "Crash into 2" receive different weights of -1, and -2 respectively. This highlights that the decision to switch is more penalized when the number of individuals on track is greater. In Fig.7, as denoted by the vertical lines, the human preferred credence values indicate that the decisions of the agent using these credence always prefer to switch when the number of individuals on track is greater than three. This contrasts with the decisions made with random credence where the agents chose to do nothing for slightly more than half of the cases. This decision contrasts with the decision observed in the classic scenario where the presence of three individuals warranted a decision to switch. This also depicts an important finding that the classic scenario cannot be generalized to other scenarios without considering the specifics of those scenarios.

*b) Variance Voting:* Fig. 8(a) and Fig. 8(b) represent the behavior of the model for the modified double scenario under the variance voting mechanism using random credence and using our approach respectively. Fig. 8(b) also depicts the difference in decision-making when comparing against 1 v X scenarios, highlighting how changing the number of individuals on alternate path of the track changes the agent's decision. This can also be explained by the difference in weights assigned to different actions by the ethical principles, already described under Nash voting. The decision-making in variance voting follows a similar pattern to Nash voting, and shows that when the number of individuals on track is greater than 2, the agent's decision guided by human preferred credence is to "switch". This decision supports the utilitarianism principle where reducing overall harm is a priority. However, this decision significantly contrasts with the decision made using random credence, where the preference to do nothing is predominantly observed as depicted in Fig. 8(a). This shows that relying on agents with random credence for decision-making often leads to choices in decisions that do not align with human preferences for the majority of the scenarios.

## V. Conclusion

The absence of moral ground truth for decisions of AVs is a major hindrance to ethical decision-making. This paper alleviates this issue by introducing the concept of human preferences for AVs by taking human-preferred empirical data to generate credence towards a particular theory. Our approach provides a mathematical rationale behind the decisions guided by credence generated through human preferences, adding a layer of explainability to the decisions. It was found that decisions made by an RL agent can be significantly improved to align their decisions with those made by humans, making them viable for practical applications. This practicability can

further be enhanced by integrating additional ethical theories such as Justice Theory to introduce the concept of "fairness" into the decision-making capability of AVs.

## VI. Future Work

The "Moral Machine Dataset" we used for generating the credence values didn't consider all the different scenarios that could affect the decision making process. These scenarios do not account for varying degrees of risk involved, vehicle speeds, or environmental conditions such as weather or road quality. These factors could significantly influence both human preferences and optimal decision-making in real-world situations. We would incorporate more variables to provide a more comprehensive solution for ethical decision-making in AVs leveraging more human-preference datasets [18], [19]. We can also utilize the National Highway Traffic Safety Administration (NHTSA) dataset [17], which provides detailed information about crash conditions, including vehicle speeds and collision information.

Furthermore, another future work could be to deal with the inherent biases in ethical theories. Deontological ethics focuses on the intrinsic rightness or wrongness of actions rather than their consequences. This could lead to a bias towards inaction in certain scenarios that we considered. Deontological theory does not allow us to consider people as a means to an end which could result in decisions that prioritize avoiding direct harm over taking action to minimize overall harm. Future iterations of the model could explore ways to balance this potential bias, by introducing the weighting of deontological principles or by incorporating additional ethical frameworks.

## References


[1] "Autonomous Vehicle Market Size, Share, Trends, Report 2022-2030," [online] Available: https://www.precedenceresearch.com/autonomous-vehicle-market, accessed on: Apr. 7, 2024.

[2] B. Meder et al., "How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty," *Risk analysis*, vol. 39, no. 2, pp. 295–314, Feb 2019.

[3] J. Z. Leibo et al., "Multi-agent Reinforcement Learning in Sequential Social Dilemmas," Available: https://arxiv.org/pdf/1702.03037.pdf, 2017.

[4] D. Abel, J. MacGlashan, and M. L. Littman, "Reinforcement learning as a framework for ethical decision making", in proc. *Workshops at the thirtieth AAAI conference on artificial intelligence*, Mar 2016.

[5] A. Ecoffet, J. Lehman, "Reinforcement Learning Under Moral Uncertainty", [Online]. Available: http://proceedings.mlr.press/v139/ecoffet21a/ecoffet21a.pdf, accessed on: Apr. 7, 2024.

[6] A. Loreggia et. al., "Making human-like moral decisions", *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 447–454, Jul 2022.

[7] R. A. Bradley, M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons", *Biometrika*, vol. 39, no. 3, pp. 324, Dec. 1952.

[8] E. Awad et al., "The Moral Machine experiment", *Nature*, vol. 563, no. 7729, pp. 59–64, doi: https://doi.org/10.1038/s41586-018-0637-6, Oct. 2018.

[9] Uber Research, [online] Avaliable: https://github.com/uber-research/normative-uncertainty, accessed on Apr. 07, 2024.

[10] T. Gill, "Ethical dilemmas are really important to potential adopters of autonomous vehicles", *Ethics and information technology*, vol. 23, no. 4, pp. 657-673, 2021.

[11] S. M. Thornton, S. Pan, S. M. Erlien and J. C. Gerdes, "Incorporating Ethical Considerations Into Automated Vehicle Control," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1429-1439, doi: 10.1109/TITS.2016.2609339, June 2017.

[12] K. Bogosian, "Implementation of Moral Uncertainty in Intelligent Machines," *Minds & Machines*, vol. 27, pp. 591-608, 2017

[13] W. MacAskill, "Normative Uncertainty", *Oxford University, UK*, 2014.

[14] I. Gabriel, "Artificial intelligence, values, and alignment," *Minds and Machines*, vol. 30, no. 3, pp. 411-437, 2020.

[15] M. Geisslinger, F. Poszler, J. Betz, C. Lütge, and M. Lienkamp, "Autonomous driving ethics: From trolley problem to ethics of risk," *Philosophy & Technology*, vol. 34, no. 4, pp. 1033-1055, 2021.

[16] N. J. Goodall, "Away from trolley problems and toward risk management," *Applied Artificial Intelligence*, vol. 30, no. 8, pp. 810–821, 2016.

[17] National Highway Traffic Safety Administration, "Fatality Analysis Reporting System (FARS)", [Online]. Available: [https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars], accessed on August. 25, 2024.

[18] B. Meder, N. Fleischhut, N.-C. Krumnau, and M. R. Waldmann, "How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty," *Risk Analysis*, vol. 39, no. 2, pp. 295–314, 2019.

[19] S. Krüger and M. Uhl, "Autonomous vehicles and moral judgments under risk," *Transportation Research Part A: Policy and Practice*, vol. 155, pp. 1–10, 2022.