

Exploring the Mechanism of Moral Machine Experiment Review

Yukai Qiao*

Department of Psychology, New York University Shanghai, Yang Si West Road, Shanghai, China

*Corresponding author: yq2194@nyu.edu

Abstract. The advancement of artificial intelligence has intensified the focus on machine moral decision-making. This article delves into the Moral Machine experiment, a platform addressing ethical dilemmas self-driving cars face. This study seeks to illuminate the intricacies of the moral dilemmas faced by autonomous systems and the potential implications for the broader landscape of AI ethics. The study aims to uncover essential principles for machine ethics by exploring global perspectives on these moral quandaries. The results reveal shared preferences for sparing humans, saving more lives, and prioritizing young lives. However, cultural clusters exhibit diverse preferences, emphasizing the complexity of cross-cultural moral choices. This article analyzes personal, social, country-level, and cultural predictors contributing to these variations. Ultimately, it underscores the challenge of establishing universal moral rules for AI and encourages moral thinking in navigating real-world challenges.

Keywords: Moral Psychology, AI Morality, Self-driving Vehicles, Trolley Problem.

1. Introduction

In the era of rapid advancements in artificial intelligence, the ethical dimensions of machine decision-making have emerged as a pivotal concern. As AI technologies, such as self-driving cars, become increasingly integrated into our lives, questions surrounding how these machines should navigate morally complex situations have captured the attention of researchers, ethicists, and society at large. The ability of machines to make choices that reflect ethical considerations, akin to human decision-making, has prompted extensive exploration into the field of machine moral decision-making. A compelling illustration of this ethical problem resides in the context of autonomous vehicles. Within this sphere, a poignant moral dilemma arises when considering situations where an unavoidable accident is imminent, and a choice must be made regarding preserving life. Imagine a self-driving car facing a scenario where it must decide whether to swerve and endanger its passengers to save pedestrians or prioritize its passengers' safety at the expense of pedestrians. This scenario encapsulates the intricate balance between human life, technological capabilities, and moral principles that have thrust machine moral decision-making into the forefront of societal discourse. The fundamental purpose of this research is to delve into the realm of machine moral decision-making comprehensively. By examining the nuances of the moral machine experiment, we aim to decipher the underpinnings of human moral choices and their translation into machine-based decisions. This study seeks to illuminate the intricacies of the moral dilemmas faced by autonomous systems and the potential implications for the broader landscape of AI ethics. Through rigorous analysis and thoughtful consideration, we endeavor to contribute to the ongoing dialogue surrounding the ethical integration of artificial intelligence into our lives.

2. Literature review

The exploration of literature within the context of machine moral decision-making plays a pivotal role in establishing a crucial groundwork for comprehending the existing reservoir of knowledge, methodologies, and insights that shape the trajectory of the moral machine experiment. This section immerses itself in the fundamental concepts and discoveries extracted from diverse studies, furnishing a comprehensive amalgamation of pertinent scholarly works.

The primary resource that forms the crux of our study is the Moral Machine experiment conducted by Awad et al. [1]. In their scholarly discourse, the researchers delve extensively into global moral

inclinations. This inquiry spans multiple dimensions: firstly, the researchers chronicle the pervasive global preferences; secondly, they meticulously document the intricacies of individual variations in preferences, contingent upon the demographics of respondents. The third facet unfolds as a revelation of cross-cultural disparities in ethical perspectives, leading to the discernment of three distinct clusters of countries. Remarkably, these disparities are closely intertwined with contemporary institutional constructs and deeply entrenched cultural traits. Conclusively, the researchers discuss the potential for these divergent preferences to contribute to developing universally accepted principles governing machine ethics. Notably, the data employed in this article is accessible to the public.

Firmly rooted in meticulously designed experiments and the active participation of individuals spanning the globe, the moral machine experiment has unearthed a cascade of invaluable and thought-provoking experimental outcomes. Foremost among these outcomes are the three preferences that stand as shared bedrocks worldwide: the pronounced inclination to prioritize humans over animals, the endorsement of saving more lives, and the penchant for safeguarding the young. These preferences, representing the cornerstone of ethical considerations, bear the potential to serve as fundamental building blocks for the establishment of ethical frameworks for machines.

Expanding their purview, the experimenters partitioned countries into three distinct cultural clusters. The Western cluster encompasses North America and numerous European countries, predominantly characterized by Catholic, Orthodox, and Christian cultural affiliations. The Eastern cluster encompasses far eastern countries like Japan and Korea, tied to the Confucianist cultural group, alongside Islamic nations such as Indonesia, Pakistan, and Saudi Arabia. Lastly, the Southern cluster embraces Latin American countries of Central and South America. This segmentation offers captivating insights, as participants from distinct cultural clusters exhibit a rich tapestry of diverse preferences.

In the Western cluster, there is a prevailing inclination to spare more lives at the expense of animals, alongside a preference to spare women and endorse inaction. Conversely, the Eastern cluster prioritizes saving older individuals and pedestrians yet showcases a weaker inclination to spare women. The Southern cluster emerges with a robust preference for sparing women and young people, accompanied by a strong bias towards sparing fit individuals, juxtaposed against a comparably weaker preference for sparing humans over pets.

In this multifaceted exploration of global moral inclinations and their intersections with culture, demographics, and ethics, the moral machine experiment resounds as a beacon, beckoning us to traverse the convoluted contours of human values, ethics, and technology. The diverse preferences from this experiment reverberate far beyond its confines, inviting collective contemplation, discourse, and action toward the ethical integration of artificial intelligence into our collective reality.

2.1. Machine Moral Decision-Making: Definition and Implications

Machine moral decision-making is a burgeoning field that grapples with the complex challenge of imbuing artificial intelligence systems with the capacity to navigate ethical dilemmas and make morally sound choices. At its core, this concept revolves around endowing machines with the ability to analyze situations, weigh various moral considerations, and select actions that align with established ethical principles, often paralleling human decision-making processes.

The essence of machine moral decision-making resides in integrating two critical components: computational algorithms and ethical frameworks. Computational algorithms enable machines to process information, assess potential outcomes, and generate possible courses of action. Concurrently, ethical frameworks provide the foundation for evaluating choices by incorporating established moral principles, cultural norms, and societal values.

In practical terms, machine moral decision-making is applied in many domains, from autonomous vehicles to medical diagnostics. In the context of autonomous vehicles, the challenge lies in programming these vehicles to make instantaneous choices that prioritize human safety while minimizing harm to all parties involved. This poses intricate questions about the valuation of different

lives, the role of utilitarianism versus deontological ethics, and the extent to which AI systems can replicate human moral intuition.

The implications of successful machine moral decision-making are far-reaching. They extend beyond technological advancements to touch the ethical fabric of our society. As AI systems become more capable of emulating human moral reasoning, the conversation inevitably shifts to concerns about accountability, transparency, and the establishment of universally accepted ethical standards for AI decision-making. Striking a balance between ethical consistency and cultural context emerges as a pivotal challenge, demanding collaborative efforts from ethicists, technologists, and policymakers.

In the subsequent sections, we delve deeper into the landscape of machine moral decision-making, examining experimental designs, anticipated results, and the multifaceted predictors that shape individuals' responses to moral dilemmas. Through this exploration, we endeavor to shed light on the intricate interplay between AI technology, human morality, and the ever-evolving ethical contours of the modern world [2].

2.2. Expected Experimental Design: Literature Compilation

A comprehensive compilation of relevant literature serves as a cornerstone to delineate the landscape of machine moral decision-making. Drawing from diverse studies, research papers, and experiments, this section seeks to distill key insights, methodologies, and experimental paradigms that have shaped the anticipated design of the moral machine experiment.

The design of the experiment is intricately woven to unveil individuals' moral choices when confronted with complex scenarios. The synthesis of existing literature informs the creation of multifaceted ethical dilemmas, ensuring the experiment captures a broad spectrum of moral considerations. By amalgamating insights from disciplines such as psychology, ethics, and human-computer interaction, the experimental design endeavors to simulate real-world moral quandaries encountered by autonomous vehicles.

2.3. Expected Experimental Results

Envisioned as a crucible of insights into human moral decision-making, the moral machine experiment is poised to yield a mosaic of anticipated results. Rooted in participants' diverse preferences and perspectives, the experiment is anticipated to unearth commonalities and disparities in moral choices across cultural clusters, gender, and belief systems.

The following expected result delves into how one's cultural background influences one's moral principles. Following the confirmation of the vital link between morality and culture in Edmond Awad's research [3], The findings reveal that there isn't a substantial distinction in moral decision-making between individuals who hold religious beliefs and those who do not. However, it is noteworthy that individuals with religious convictions tend to be more inclined to prioritize the protection of older adults over wealthier individuals. The available literature indicates that Eastern cultures exhibit lower levels of ageism than Western cultures. This difference is attributed to the cultural norms in the East, which emphasize a greater reverence for older individuals. This may imply that in the trolley problem, individuals affected by Eastern culture clusters are more inclined to preserve the lives of older people than Western culture clusters.

The study done by Atari et al. suggested that women scored higher than men on Care, Fairness, and Purity across cultures. The sex differences in these moral judgments seem replicable and robust across cultures [2]. A Study done by Herzog shows that women, on average, show higher levels of positive behaviors and attitudes toward animals. In contrast, men typically have higher levels of negative attitudes and behaviors.

Another study done by Cundiff et al. supports this argument. Gender differences in perceptions suggest that, relative to males, females are more likely to report higher levels of ethnic or cultural empathy and also have a higher preference for women in leadership and authority positions [4]

Besides, a Study by Faulhaber et al shows that As a common pattern, most participants consistently made choices following a practical approach, placing paramount importance on saving the highest possible number of avatars while showing minimal susceptibility to external factors. These findings, following the practical viewpoint regarding moral decision-making, provide substantiation for advocating the imposition of obligatory ethical guidelines in the domain of autonomous driving vehicles [5]

The earlier paragraphs' findings support the observations made in the third paragraph. The results suggest that women consistently score higher than men in dimensions related to Care, Fairness, and Purity, irrespective of cultural backgrounds. This consistency in gender-based moral judgments aligns with the assertion that women often receive more support historically and that their hardships are regarded as more profound. It also corresponds to the notion that losing an innocent woman is considered more distressing than a man's.

Furthermore, the data indicates that gender differences extend to perceptions, with females tending to report higher levels of ethnic or cultural empathy and a stronger preference for women in leadership and authority positions. These variations in gender perceptions support the argument that decisions, especially those involving elderly individuals, often exhibit utilitarian reasoning that prioritizes the protection of women and children.

In parallel, as highlighted in the third paragraph, the general trend among participants is to make practical decisions, prioritizing preserving the maximum number of avatars. This inclination towards utilitarian decision-making reinforces the notion that decisions, even in the "moral machine" experiment, prioritize saving lives, especially those of women and children, over other considerations.

According to the research done by Husted, the link between moral judgment and behavior is related to individualism and collectivism [6]. The predicted experimental results are likely to be: In ethical problems, those with individualistic leanings generally oppose making decisions on behalf of others. Nonetheless, situations like those encountered in the "moral machine" occasionally demand such choices, underscoring the intrinsic worth of each individual. Individualists favor decisions maximizing the number of lives saved in scenarios where quantity is crucial. Additionally, they prefer preserving the lives of younger individuals, emphasizing their potential and contributions to society. In contrast, collectivist societies, which hold elders in high regard, are less inclined to prioritize preserving younger members [7, 8].

Another expected result is that children and young individuals are less likely to spare older adults. Study shows that by observing others or due to in-group biases, children as early as three years old hold negative stereotypes about older people. In this case, they are more likely to spare the young, while elderlies tend to spare elderlies due to in-group biases. This indicates an age-related difference [9]

The last hypothesis is that more empathetic individuals are likelier to spare disadvantaged groups. More empathetic individuals would better consider other people's perspectives and empathetically understand the feelings and situations of disadvantaged groups. Therefore, they may be more likely to spare vulnerable groups like homeless people. Study shows there is a correlation between higher levels of empathy and a greater likelihood to spare disadvantaged or vulnerable groups in moral decision-making scenarios [10]

2.4. Analysis and Discussion

The forthcoming analysis and discussion will delve into the rich tapestry of the anticipated experimental results. By probing the factors contributing to variations in moral preferences, this section seeks to unravel the diverse influences of personal, social, and cultural predictors. Exploring why participants from different regions and cultural clusters exhibit disparate preferences for the same moral dilemmas will offer profound insights into the complexities of human moral reasoning.

Through a holistic lens, this analysis underscores the dynamic interplay between individual autonomy and collective values, illuminating the role of cultural and societal contexts in moral

decision-making. Additionally, it scrutinizes the potential implications of these results for the field of machine ethics and the broader integration of AI into our lives.

2.5. Research Insights: Implications for Reality

This study embarks on a journey to decipher the intricate mosaic of human moral choices and aspires to offer valuable insights for real-world applications. This research endeavors to illuminate pathways toward ethically sound AI decision-making by unraveling the nuanced predictors that shape moral preferences. The experiment's implications cascade into the domain of autonomous systems, encouraging deliberations on establishing universal moral guidelines for AI algorithms.

As this research unearths the profound impact of culture, society, and personal beliefs on moral choices, it underscores the necessity for interdisciplinary collaborations among technologists, ethicists, and policymakers. By anchoring our analysis in empirical evidence and ethical reflection, we strive to contribute to a more comprehensive understanding of the intricate interplay between machine moral decision-making and its implications for navigating the evolving landscape of artificial intelligence in our reality.

3. Conclusion

Males In the wake of rapid advancements in artificial intelligence, the intricate landscape of machine moral decision-making has come into sharp focus. This journey through the moral machine experiment has illuminated the multifaceted nature of human ethical choices and the endeavor to transpose these nuances onto the canvas of AI systems. As we draw the curtains on this exploration, several resounding conclusions emerge, underscoring the complex interplay of culture, society, and individual beliefs in shaping the ethical fabric of our technological landscape.

The experiment's findings vividly portray the global moral preferences that unite and differentiate us. The resounding echo of sparing humans over animals, prioritizing more lives saved, and safeguarding the young resonates across cultures, underscoring the foundational pillars upon which machine ethics can potentially be constructed. These shared preferences emerge as integral components for establishing a common framework that aligns with the moral compass of humanity.

However, the experiment's canvas also boasts vibrant strokes of cultural variation. The emergence of Western, Eastern, and Southern cultural clusters reveals the profound impact of societal norms and historical legacies on ethical choices. These clusters serve as a testament to the mosaic of human perspectives, each influenced by unique narratives that contribute to the tapestry of our global society.

The revelation of correlations between these cultural clusters, modern institutions, and deeply ingrained cultural traits offers a panoramic view of the intricate forces that shape our moral compass. The analysis of personal, social, and country-level predictors unearths the delicate balance between individual autonomy and collective values, challenging the notion of universal moral principles. This interplay between the individual and the collective accentuates the complexity of formulating ethically grounded guidelines for AI decision-making.

In this rapidly changing landscape, pursuing global, socially acceptable principles for machine ethics remains a formidable challenge. The experiment's curtain call echoes the words of philosopher Ludwig Wittgenstein: "The limits of my language mean the limits of my world." Similarly, the experiment's outcomes remind us that the limits of our ethics may shape the boundaries of our AI-powered reality. The journey continues, inviting society to contemplate the universal values underpinning our collective moral tapestry. As the grand symphony of AI ethics unfolds, we must continue to engage, reflect, and evolve, ensuring that the algorithms we create mirror our intelligence and humanity.

References

- [1] Awad, E., Dsouza, S., Kim, R. et al. The Moral Machine experiment. *Nature*, 2018, 563, 59-64.

- [2] Atari, M., Lai, M. H., Dehghani, M. Sex differences in moral judgments across 67 countries. The Royal Society Publishing, 2020.
- [3] 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 of the United States of America*, 2020, 117(5), 2332-2337.
- [4] Cundiff, N. L., Komarraju, M. Gender Differences in Ethnocultural Empathy and Attitudes Toward Men and Women in Authority. *Journal of Leadership and Organizational Studies*, 2008, 15(1), 5-15.
- [5] Faulhaber, A.K., Dittmer, A., Blind, F. Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles. *Sci Eng Ethics*, 2019, 25, 399-418.
- [6] Husted, B.W., Allen, D.B. Toward a Model of Cross-Cultural Business Ethics: The Impact of Individualism and Collectivism on the Ethical Decision-Making Process. *J Bus Ethics*, 2008, 82, 293-305
- [7] Herzog, H. A. Gender Differences in Human–Animal Interactions: A Review. Paws.wcu Press, 2007.
- [8] Nelson, T. D. Ageism In Nelson T. D. (Ed.), *Handbook of prejudice, stereotyping, and discrimination* New York, NY: Psychology Press, 2009.
- [9] Decety, J., Wheatley, T. *The Moral Brain: A Multidisciplinary Perspective*. The MIT Press, 2015.
- [10] Mendonca, J., Marques, S., & Abrams, D. *Children’s Attitudes toward Older People: Current and Future Directions*. Springer Press, 2008.