When I Choose for You: Moral Reasoning in Third-Person Ethics for HRI

Hannah Wiedemann[1,2], Serena Marchesi[1] and Agnieszka Wykowska[1]

[1]S4HRI – Istituto Italiano di Tecnologia, Genova, Italy
[2]Department of Neuroscience and Rehabilitation, University of Ferrara, Ferrara, Italy

Abstract

As robots assume morally sensitive roles in our environments, understanding how humans make moral decisions for others becomes crucial. This study investigates third-person moral reasoning using dilemmas where participants chose between deontological and utilitarian actions on behalf of a confederate, mirroring real-world scenarios such as medical contexts, where decisions can be made on behalf of someone else. Results show that personal relationship dilemmas prompted more rule-based choices, while life-and-death scenarios led to longer response times, suggesting greater cognitive conflict. These findings offer a baseline for future comparisons with robot-directed decisions and shed light on how people navigate ethical choices in complex social settings.

Keywords

Moral Decision Making, Human-Robot Interaction, Value Projection

1. Introduction

Moral decision-making is a complex cognitive process that draws upon fundamental psychological functions, including perception, memory, and attention [1]. These capabilities enable individuals to interpret social scenarios, retrieve relevant experiences, and identify critical information, all of which are essential for evaluating the moral appropriateness of a given action [2] [3]. In real-world situations, especially when individuals are confronted with moral dilemmas, these processes work in tandem to guide judgments about what is right or wrong; and shape subsequent behaviour (for a review see [4]).

As intelligent systems, including robots, become increasingly integrated into our daily lives — from autonomous vehicles to digital assistants — they inevitably become part of our social environments and moral implications [5]. This integration raises important questions about how we make moral decisions involving artificial agents, and whether our reasoning differs depending on whether the agent is human or machine [6]. In many real-life situations—such as a doctor deciding on a treatment plan for an unconscious patient—people must make moral decisions on behalf of others. Understanding human moral decision-making in these contexts is therefore essential for building ethically aligned and socially acceptable technologies [7] [8].

One influential model of moral judgment is Greene's [9] dual-process theory, which posits that two distinct cognitive systems, emotional and rational, underlie moral reasoning. Emotional responses are typically fast, automatic, and associated with deontological judgments, where the morality of an action is judged based on rules or duties. In contrast, rational processing is slower, deliberative, and linked to utilitarian judgments, where moral decisions are based on outcomes and the greater good [10]. This theory highlights how competing cognitive systems can lead to different moral conclusions depending on which system is dominant in a given context.

Importantly, different moral decision-making contexts, such as issues concerning privacy, health, or harm, can significantly influence how individuals arrive at their moral judgments [11]. For example, people may rely more on emotional reasoning in high-stakes, personal scenarios, whereas impersonal

2025 Workshop on Awareness in Learning Agents
hannah.wiedemann@iit.it (H. Wiedemann); serena.marchesi@iit.it (S. Marchesi); agnieszka.wykowska@iit.it (A. Wykowska)
0009–0008–2506–2941 (H. Wiedemann); 0000–0001–9931–156X (S. Marchesi); 0000–0003–3323–7357 (A. Wykowska)

or abstract dilemmas may elicit more utilitarian reasoning [12]. These contextual variations must be taken into account to fully understand how moral decisions are made.

Moral reasoning does not occur in a vacuum. It is shaped by the social and situational context in which decisions are made. Recent advances in human-robot interaction (HRI) raise critical questions about how humans apply moral principles when interacting with artificial agents [13]. While extensive research has explored moral judgment in human-human contexts, far less is known about how people make moral decisions when robots are involved [14]. Specifically, it remains unclear whether humans apply the same moral frameworks to robots as they do to other humans.

1.1. Aim

To understand moral decision-making in human–robot interaction, it is crucial to first examine how people make such decisions for other humans. We hypothesize that participants will make more deontological judgments when deciding on behalf of another human, driven by emotional and intuitive processing. This aligns with dual-process theories of moral cognition [9]. In contrast, in future experiments where the human agent is replaced by a robot—devoid of human-like emotional cues, participants are expected to favor more utilitarian decisions, as the removal of emotional engagement may diminish deontological inclinations. Empirical work supports this prediction: people tend to expect robots to make utilitarian choices (e.g., sacrificing one to save many) more than humans do, and view such decisions as more permissible when performed by robots [15]. While our broader project will directly compare responses to human versus robot agents, this study represents the initial, foundational step—establishing a baseline of human moral judgment. By focusing on decisions made for another human, we lay the groundwork for future comparisons with robotic agents, which will allow us to interpret how moral reasoning (utilitarian vs. deontological responses) may differ in the context of robots.

2. Methods

2.1. Participants

Twenty-one participants (11 female; age 19–65, M = 32.00 SD = 13.40) were recruited, with a final sample of 20 (10 female; M = 31.10, SD = 13.73) after excluding one due to withdrawal. This represents a preliminary dataset and further recruitment is ongoing. All had normal or corrected vision, no neurological conditions, were naïve to the study's purpose, and fluent in Italian. The study was approved by the Comitato Etico Regione Liguria. Participants gave written informed consent, were debriefed post-experiment, and received €10 compensation.

2.2. Experimental Design

In this study, participants completed a moral decision-making task simulating third-person ethical reasoning. In each trial, they read a moral dilemma involving a confederate who asked, "What should I do?" Participants chose between a deontological or utilitarian response on the confederate's behalf. The main manipulation was the type of dilemma, with each scenario presented once. Future studies will vary the agent (human vs. robot).

2.3. Apparatus and Stimuli

As shown in Figure 1, participants sat next to a research assistant (confederate). They viewed the experimental procedure on a 27-inch Dell monitor (120 Hz, 2560×1440), while the confederate used a 15.6-inch Dell laptop (60 Hz, 1920×1080) running the experiment. The procedure was programmed in PsychoPy 2022.2.4 [16].

Figure 1: Set up of the experiment: Confederate (left) viewing the laptop and participant (right) viewing the monitor and using the labeled keyboard for answer input.

2.4. Scenarios and Validation

The scenarios and response options were generated using ChatGPT 4o [17]. 50 scenarios were then validated through an online questionnaire through Prolific. The validation process is described in [18]. After validation, the total amount of validated scenarios and answer options were 42 (See Table 1 for an example). Furthermore, the dilemmas were categorized into seven distinct groups for future analysis.

Table 1
Example scenario and both answer options

Scenario	Deontological Answer	Utilitarian Answer
A terminally ill patient asks if their condition is improving. The truth is that they only have a few weeks left to live, and telling them might cause immense distress. However, lying might provide them comfort in their final days.	Tell the truth	Lie to the patient

2.5. Experimental Setup

The confederate introduced himself and explained the experiment.

Participants completed 42 randomly presented trials. Each trial began with a scenario displayed on screen until the participant pressed the space bar. A fixation cross (1000 ms) and a beep (500 ms)

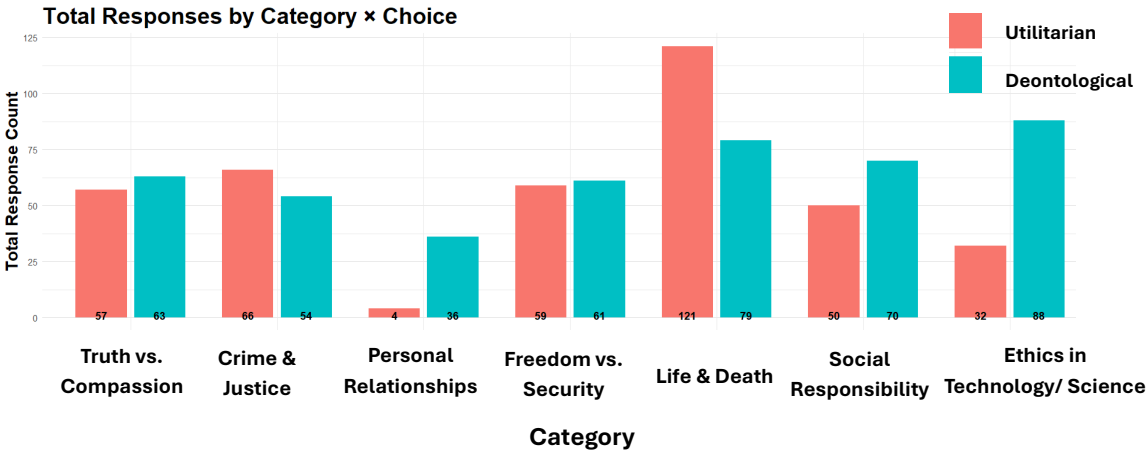

Figure 2: Distribution of Utilitarian and Deontological Responses Across Categories

preceded the confederate's prompt—"What should I do?"—shown for 3000 ms. Two response options then appeared (left/right, counterbalanced). Participants selected their choice using the 'A' (left) or 'L' (right) key. After responding, a fixation cross appeared for 1000 ms before the next trial began.

2.6. Statistical Analysis

Statistical data analysis was conducted using custom-made scripts in R Studio (version 2022.07.1) [19] [20].

Generalized Linear Mixed-Effects Models (GLMMs, glmer function) and Linear Mixed-Effects Models (LMMs, lmer function) were computed using the lme4 package [19]. The lmerTest package [21] was applied to obtain p-values for fixed effects in the LMMs. Parameters estimated (β) for fixed effects and their associated t-tests (t, p-value) were calculated using the Satterthwaite approximation method for degrees of freedom. For visualisation of results, the ggplot2 [22] and ggeffects [23] packages were used.

3. Results

3.1. Choice Outcome

Raw data visualization is presented in Figure 2.

We investigated whether different types of moral dilemmas systematically influenced individuals' choices by applying a generalized linear mixed-effects model. The binary outcome variable, *choice* (utilitarian vs. deontological), was modeled as a function of the fixed effect *category* (e.g., life and death, crime and justice), with random intercepts for both participants and dilemmas to account for repeated measures and clustering. The model employed a binomial logistic link function and was estimated using maximum likelihood with a Laplace approximation.

Including *category* as a fixed effect allowed us to test whether responses to particular categories of dilemmas differed from the designated reference category (*Truth vs Compassion*). Random intercepts for participants and dilemmas controlled for individual- and scenario-specific variability.

The fixed effects analysis indicated that most categories did not differ significantly from the reference category. However, the *Personal Relationships* category was associated with a significantly higher log-odds of making a deontological choice relative to *Truth vs Compassion* (estimate = 2.80, SE = 1.34, z = 2.08, p = .037). No other categories showed statistically significant differences from the reference. A full set of model coefficients is provided in the supplementary materials.

The model intercept (estimate = 0.107, p = .863) represents the estimated log-odds of a deontological choice in the reference category (*Truth vs Compassion*). Thus, the intercept should be interpreted narrowly as the expected value for that category, rather than as a general or overall baseline.

Regarding random effects, substantial variability was observed across dilemmas (variance = 1.943, SD = 1.394), indicating that the specific scenario influenced responses strongly. Participant-level variance was smaller (0.285, SD = 0.534) but still indicated individual differences in choice tendencies.

3.2. RT Outcome

A linear mixed-effects model was used to investigate the effects of *category* and *choice* on log-transformed response time. The model, estimated using restricted maximum likelihood, included random intercepts for both *participant* and *dilemma*.

The fixed effects analysis suggested that most categories did not differ significantly from the reference (*Truth vs Compassion*). However, the *Life and Death* category was associated with longer response times relative to the reference (estimate = 0.362, p = .028; Overall M= 5.24, SD= 6.25; Deontological M= 6.63, SD= 7.19; Utilitarian M= 4.34, SD= 5.39). The *Social Responsibility* category showed a marginal trend toward slower responses (estimate = 0.304, p = .094), although this effect was not statistically significant.

The effect of *choice* was positive but non-significant (estimate = 0.069, p = .139), indicating that deontological vs. utilitarian responding did not, in itself, reliably affect response time.

The model intercept (estimate = 0.922, $p < .001$) reflects the estimated mean log response time in the reference condition (*Truth vs Compassion*), rather than a general baseline across all categories.

With respect to random effects, the variance associated with *dilemmas* (0.0787) was greater than that associated with *participants* (0.0431), suggesting that scenario-level differences contributed more strongly to variation in response times than individual differences. The residual variance was 0.3051, reflecting considerable within-condition variability in response times.

4. Discussion and Conclusion

This study aimed to establish a baseline for understanding how people make moral decisions on behalf of others in human-agent scenarios. Our findings highlight the complexity and variability of such judgments, particularly across different types of moral dilemmas.

We found that dilemmas involving personal relationships prompted more deontological responses, suggesting that close interpersonal bonds increase rule based reasoning—likely due to stronger emotional engagement. In contrast, other categories did not predict moral choices consistently, indicating more mixed or conflicted reasoning.

Life and Death dilemmas did not shift moral choices significantly but were associated with longer response times, suggesting greater cognitive conflict and engagement of slower, more reflective processing, in line with Greene's dualprocess model.

These findings underscore the complexity of moral reasoning when individuals make decisions for others. The fact that some dilemma types affect moral preferences while others do not suggests the need for a more detailed understanding of the contextual, emotional, and interpersonal factors that shape moral judgment.

4.1. Future Directions

This study establishes a baseline for comparing human-to-human moral decisions with those involving robot agents. Future work will explore whether the presence of a robot alters moral reasoning or shifts the balance between deontological and utilitarian choices. These insights will inform how people adapt ethical frameworks in technologically mediated settings and guide the development of ethically aligned AI systems.

Acknowledgments

This work has received support from the European Union under the European Innovation Council (EIC) research and innovation programme, Project "VaLue-aware AI (VALAWAI)", Grant Agreement number 101070930.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, in order to: Create stimuli, Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Garrigan, A. L. Adlam, P. E. Langdon, Moral decision-making and moral development: Toward an integrative framework, Developmental Review 49 (2018) 80–100. URL: https://doi.org/10.1016/j.dr.2018.06.001. doi:10.1016/j.dr.2018.06.001.

[2] C. D. Frith, T. Singer, The role of social cognition in decision making, Philosophical Transactions of the Royal Society B Biological Sciences 363 (2008) 3875–3886. URL: https://doi.org/10.1098/rstb.2008.0156. doi:10.1098/rstb.2008.0156.

[3] N. Kar, B. Kar, Social cognition and individual effectiveness in interpersonal scenarios: A conceptual review, Journal of Mental Health and Human Behaviour 22 (2017) 27–34. URL: https://doi.org/10.4103/0971-8990.210705. doi:10.4103/0971-8990.210705.

[4] S. Guglielmo, Moral judgment as information processing: An integrative review, Frontiers in Psychology 6 (2015) 1637. URL: https://doi.org/10.3389/fpsyg.2015.01637. doi:10.3389/fpsyg.2015.01637.

[5] T. J. Prescott, J. M. Robillard, Are friends electric? The benefits and risks of human-robot relationships, iScience 24 (2020) 101993. URL: https://doi.org/10.1016/j.isci.2020.101993. doi:10.1016/j.isci.2020.101993.

[6] Z. O'Reilly, S. Marchesi, A. Wykowska, The impact of action descriptions on attribution of moral responsibility towards robots, Scientific Reports 15 (2025) 4128. URL: https://doi.org/10.1038/s41598-024-79027-5. doi:10.1038/s41598-024-79027-5.

[7] P. Boddington, The rise of AI ethics, in: AI Ethics, Artificial Intelligence: Foundations, Theory, and Algorithms, Springer, Singapore, 2023. URL: https://doi.org/10.1007/978-981-19-9382-4_2. doi:10.1007/978-981-19-9382-4_2.

[8] S. Marchesi, A. Wykowska, Designing robots that are accepted in human social environments: Anthropomorphism, the intentional stance, cultural norms and values, and societal implications, in: L. Fortunati, A. Edwards (Eds.), The De Gruyter Handbook of Robots in Society and Culture, De Gruyter, Berlin, Boston, 2024, pp. 63–84. URL: https://doi.org/10.1515/9783110792270-004. doi:10.1515/9783110792270-004.

[9] J. D. Greene, The secret joke of kant's soul, in: W. Sinnott-Armstrong (Ed.), Moral Psychology, Volume 3: The Neuroscience of Morality, MIT Press, 2001, pp. 35–79.

[10] K. K. Denton, D. L. Krebs, Rational and emotional sources of moral decision-making: An evolutionary-developmental account, Evolutionary Psychological Science 3 (2016) 72–85. URL: https://doi.org/10.1007/s40806-016-0067-3. doi:10.1007/s40806-016-0067-3.

[11] K. Ciupinska, S. Marchesi, G. A. Abbo, T. Belpaeme, A. Wykowska, Awareprompt: Using diffusion models to create methods for measuring value-aware ai architectures., in: ICAART (3), 2024, pp. 1436–1443.

[12] E. M. Polo, A. Farabbi, M. Mollura, L. Mainardi, R. Barbieri, Understanding the role of emotion in decision making process: Using machine learning to analyze physiological responses to visual, auditory, and combined stimulation, Frontiers in Human Neuroscience 17 (2024). URL: https://doi.org/10.3389/fnhum.2023.1286621. doi:10.3389/fnhum.2023.1286621.

[13] S. Marchesi, K. Ciupinska, D. De Tommaso, A. Wykowska, Value projection. humans perception of a humanoid robot in moral contexts., 2024. URL: osf.io/preprints/psyarxiv/j6bcn_v1. doi:10.31234/osf.io/j6bcn_v1.

[14] R. Wullenkord, F. Eyssel, Societal and Ethical Issues in HRI, Current Robotics Reports 1 (2020) 85–96. URL: https://doi.org/10.1007/s43154-020-00010-9. doi:10.1007/s43154-020-00010-9.

[15] M. Lee, P. Ruijten, L. Frank, Y. de Kort, W. IJsselsteijn, People may punish, but not blame robots, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411764.3445284. doi:10.1145/3411764.3445284.

[16] J. W. Peirce, PsychoPy—psychophysics software in python, Journal of Neuroscience Methods 162 (2007) 8–13. URL: https://doi.org/10.1016/j.jneumeth.2006.11.017. doi:10.1016/j.jneumeth.2006.11.017.

[17] OpenAI, Chatgpt (version 4o) [large language model], https://chat.openai.com/, 2024. https://chat.openai.com/.

[18] H. Wiedemann, S. Marchesi, A. Wykowska, Moral dilemmas as tools to study human moral cognition in human-machine interaction, 2025. Preprint.

[19] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, 2014. URL: https://doi.org/10.48550/arXiv.1406.5823. doi:10.48550/arXiv.1406.5823. arXiv:1406.5823, arXiv preprint.

[20] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015. URL: https://www.R-project.org/.

[21] A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, lmerTest package: Tests in linear mixed effects models, Journal of Statistical Software 82 (2017) 1–26. URL: https://doi.org/10.18637/jss.v082.i13. doi:10.18637/jss.v082.i13.

[22] H. Wickham, ggplot2, Wiley Interdisciplinary Reviews: Computational Statistics 3 (2011) 180–185. doi:10.1002/wics.147.

[23] D. Lüdecke, ggeffects: Tidy data frames of marginal effects from regression models, Journal of Open Source Software 3 (2018) 772. URL: https://doi.org/10.21105/joss.00772. doi:10.21105/joss.00772.