

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Moral Explainability Across Domains: Justification Norms and the Communicability of Decision Processes

Permalink

<https://escholarship.org/uc/item/8tv8z92m>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Nguyen, Breanna K.

Sinnott-Armstrong, Walter

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Moral Explainability Across Domains: Justification Norms and the Communicability of Decision Processes

Breanna K. Nguyen & Walter Sinnott-Armstrong

Department of Philosophy, Duke University

Abstract

Moral decisions vary in the extent to which they are expected to be justified. We investigate whether normative expectations for justification influence how effectively people articulate decision strategies that others can use. In a two-phase study ($N = 573$), participants made moral decisions in either a kidney allocation domain or a personal forgiveness domain (high vs. low justification expectations), providing written explanations of their reasoning. New participants either used these explanations to predict the original decisions or made independent decisions, establishing baseline agreement. Participants perceived greater reliance on structured processes in kidney allocation, baseline agreement was higher in kidney allocation, and explanations increased prediction accuracy in both domains, with a larger effect in forgiveness. Despite these differences, matching accuracy converged across domains when explanations were provided, suggesting that moral explainability is fundamentally about communicability rather than introspective accuracy.

Keywords: moral decision-making; post hoc reasoning; justification; explainability; articulability

Introduction

Our decision-making processes vary across contexts, as do the expectations for how decisions should be justified. In some domains, choices are accepted without explanation, whereas in others, legitimacy depends on the ability to articulate clear, principled reasons. These expectations shape not only how decisions are evaluated, but also how they are expressed in language.

This contrast raises broader questions about explainability in decision-making: **How do expectations for justification shape what gets expressed in language? How faithfully does language preserve the information needed to reflect or reproduce a decision process?** In some domains, explanations may add little beyond what shared norms already constrain. In others, they may capture decision-relevant structure that would otherwise remain implicit.

Moral decisions present a particularly interesting case for this question as they often carry strong normative expectations for justification, while being driven by intuitive or affective processes that resist clear articulation. In some cases, there is an expectation that moral decisions should be informed by principles like fairness or harm aversion. In other cases, moral decisions are accepted as deeply personal or intuitive. We use “justificatory norms” to refer to domain-level expectations about whether and how decisions should be explained.

Expectations that vary in strength across moral contexts and shape both what counts as an acceptable reason and how much pressure decision-makers feel to provide one. As a result, the standards for what counts as an appropriate justification can vary systematically across moral domains.

We illustrate this contrast by comparing decisions about kidney allocation with decisions about personal forgiveness. Kidney allocation decisions are governed by formal principles and institutional norms that demand transparency, consistency, and accountability, constraining acceptable explanations to shared, articulable criteria. By contrast, personal forgiveness decisions are shaped by emotion and interpersonal context, with little normative pressure to produce explanations that others could readily evaluate or reuse. This contrast highlights how domain-specific justification norms may shape the form and effectiveness of moral explanations.

In this paper, we investigate whether the expectation to justify a moral decision influences the communicability of people’s decision strategies. Specifically, we examine whether explanations produced in different moral domains vary in how effectively they can be used by others to predict decisions. To do so, we compare kidney allocation and forgiveness as contrasting cases that differ in their normative expectations for justification. We then situate this approach within existing theories of moral decision-making, including work on moral reasoning, justification, and conscious access to decision processes. We operationalize explainability as a functional property of language: how well one person’s explanation preserves the information needed for another person to apply it. We conclude by discussing the implications of our approach and findings for theories of moral cognition and the communicative role of moral explanations.

Norms and Expectations Across Moral Domains

Expectations for justification vary across moral contexts, reflecting both the decision structure and the magnitude of consequences. As decisions become higher-stakes and more irreversible, they invite greater pressure for structured reasoning and explicit justification (Kunreuther et al., 2002; Kahn & Baron, 1995). These pressures help explain why some moral domains enforce strong justificatory norms, while others permit greater ambiguity and personal discretion.

Kidney allocation exemplifies a domain with strong justification norms. Decisions are guided by formal principles—such as justice, utility, and medical urgency—to ensure fair-

ness and efficient use of scarce resources (De Fijter, 2010). These principles are embedded in institutional policies and reflect the moral weight of determining who receives life-saving treatment. As a result, kidney allocation is a high-stakes moral context in which decisions are expected to be grounded in clear principles and explicitly justified.

Forgiveness, by contrast, can be unconditional and unpredictable, and may be granted without consideration of the offender, the severity of their offense, or the presence of an apology (Andrews, 2000; Cowley, 2010; Gismero-Gonzalez et al., 2019). While official legal pardons involve high stakes, even these do not require explanation by explicit principles. In this study, we focus specifically on personal forgiveness in interpersonal contexts, where justification norms are low and it is often viewed as a deeply personal or emotional act rather than a deliberative one. Forgiveness has been conceptualized as emotion regulation, altruism, or an expression of agreeable personality traits (McCullough, 2001; Worthington et al., 2007). The consequences of forgiveness are generally confined to the personal sphere. As a result, forgiveness constitutes a low-stakes, low-explainability moral domain in which justification is neither expected nor required.

Access to Moral Reasoning Processes

Strong justification norms in some moral domains raise a further question relevant to explainability: To what extent do people have access to the reasoning behind their moral judgments? Although our study focuses on whether explanations enable decision reproduction rather than on introspective access per se, existing theoretical frameworks help clarify what moral explanations might capture.

Dual-Process Theory posits that moral judgment involves both intuitive and deliberative processes, with different domains engaging these systems to varying degrees (Moll et al., 2002; Greene, 2007; Paxton & Greene, 2010). Haidt's (2001) Social Intuitionist Model further suggests that moral judgments are typically generated intuitively and justified post hoc in response to social demands, suggesting that explanations may reflect constructed rationalizations rather than direct access to underlying processes.

Evidence from non-moral decision-making supports the view that decisions vary in the extent to which they are deliberative and structurally articulated. When choices are multi-attribute and governed by explicit criteria, people can often articulate their reasoning accurately (Morris et al., 2023), whereas weakly structured, preference-based choices elicit confident but post hoc explanations (Johansson et al., 2005). Applied to moral decision-making, structured domains may therefore support explanations that preserve relevant information, whereas less structured domains may yield explanations that are less tied to underlying decision patterns. Instead of adjudicating between these accounts, we use these frameworks to motivate a functional evaluation of explanations, assessing their capacity to enable another person to reproduce a pattern of decisions.

The Present Study

The present study examines how normative expectations for justification shape the informativeness of moral explanations by asking whether explanations produced under different justificatory norms differ in how well they reflect decision-making behavior. We link justification norms to behavioral measures of explainability by evaluating explanations functionally—assessing whether they enable another person to predict a pattern of decisions. We predict that when people make decisions in domains with strong normative expectations for justification (e.g., kidney allocation), explanations will more faithfully reflect individuals' decisions, whereas in less constrained domains (e.g., forgiveness), explanations will be less consistently aligned with the decisions they are intended to describe. Unlike prior work that focuses on isolated judgments or endorsement of moral principles, our approach evaluates explanations by their correspondence to behavior across a sequence of dilemmas. We assess whether the explanations people provide enable others to accurately predict their pattern of decisions, allowing us to test whether justificatory norms influence how effectively decision-relevant information is conveyed in language, without assuming that accurate behavioral prediction requires accurate introspective access.

Method

We test our hypothesis by pairing participants: one group provides justifications for their moral decisions, and another either uses these justifications to predict the original decisions or makes independent decisions to establish baseline agreement (see Figure 1). Prediction accuracy serves as a behavioral measure of how well explanations convey decision-relevant information. The study was pre-registered, with materials available at: https://osf.io/mqny2/overview?view_only=14e64cbc14a94cb1a4f8306a2f68fcef

Decision Phase

In the Decision Phase, participants were randomly assigned to either the kidney allocation or forgiveness condition, read a series of moral scenarios, and made a decision for each. They then rated their reliance on a specific decision process and provided a written description of how they made their decisions, which was later used in the Matching Phase. Participants were not informed that they would be asked to explain their decisions prior to completing the decision-making task, ensuring that the explanation demand did not influence the decisions themselves.

Participants We recruited a total of 204 adult participants from Prolific to complete an online survey through Qualtrics. Of these, 8 responses were excluded due to a failed comprehension check or a nonsensical answer to the open-ended question, which resulted in a final sample of 196 participants ($M_{age} = 35.8$, $SD_{age} = 13.1$; female = 113, male = 79, other = 4; White = 58.7%, Black or African American = 24.0%,

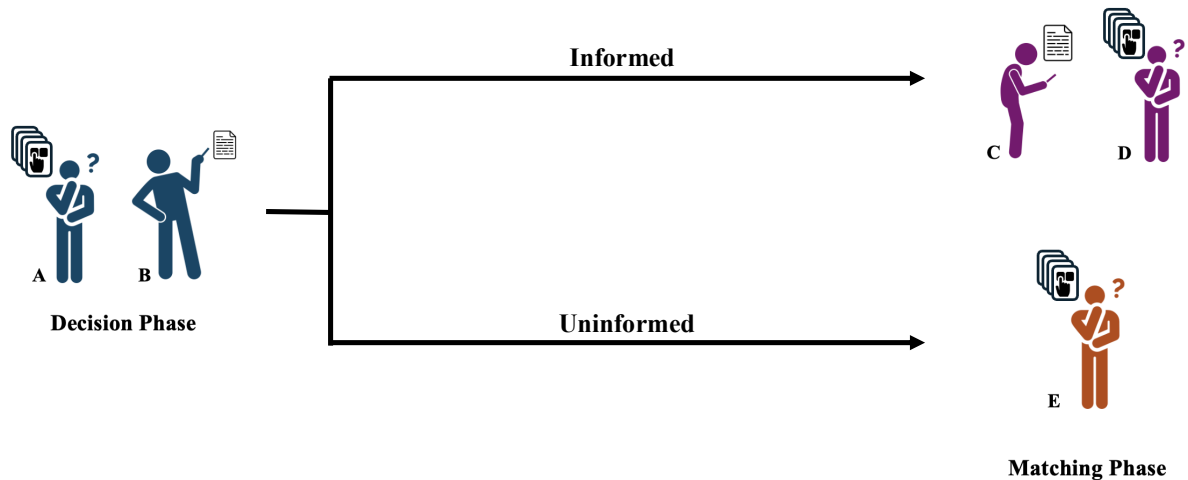


Figure 1: Our experimental design. (A) A participant makes moral decisions and (B) explains their process. (C-D) Informed participants use this explanation to predict the original decisions. (E) Uninformed participants make independent decisions on the same stimuli without knowledge of a partner, establishing baseline agreement.

Multiracial/Other = 10.2%, Asian or Asian American = 6.1%, Hispanic or Latine = 1.0%).

Kidney Allocation Condition We asked participants in the kidney allocation condition to imagine that they were part of a hospital committee tasked with deciding which of two patients (Patient A or Patient B) on the transplant list should receive a single available kidney. Across scenarios, Patient A remained constant as a reference case, while Patient B varied along multiple attributes (e.g., health severity, responsibility for illness, and emotional response), with details presented in random order. The participants' task was to decide which of the patients should receive the available kidney. Below is an example comparison:

Patient A: Jesse is a part-time librarian who generally follows doctor's advice but occasionally overlooks minor recommendations. She relies on bi-weekly treatments to manage kidney function and prevent decline. She feels a quiet determination to improve her health, acknowledging past missteps but focusing on moving forward. She has a history of mostly adhering to medical advice, with minor lapses during stressful times.

Patient B: Drew is a local mail carrier. She has a history of ignoring medical advice regarding her condition but has recently shown efforts to improve health management. She generally values health and makes an effort to maintain it, though with occasional lapses. She is experiencing life-threatening, multi-system failure due to end-stage kidney disease, requiring immediate intervention. She believes she is not at fault for her condition, insisting that external factors are entirely to blame.

Which patient would you choose to receive the kidney?

Forgiveness Condition In the forgiveness condition, participants read a series of scenarios in which someone they knew committed an interpersonal offense that might warrant forgiveness. Across scenarios, the offender varied along multiple attributes (e.g., relationship to the participant, nature of the offense, and emotional response), with details presented in random order. The participants' task was to decide whether or not they would forgive the person. Below is an example story:

Casey is your roommate. Recently, he left you to clean up after a dinner party. He admits he made a mistake but claims he was not thinking straight at the time. He has a tendency to justify or minimize his mistakes, even when confronted with evidence of harm. He is usually considerate but occasionally forgetful.

Would you forgive this person?

Decision-making Task Stimuli were constructed to maintain isomorphism, ensuring that the two domains were structurally analogous. The number of variables was identical across conditions, each variable had a direct counterpart in the other domain, and each analog variable pair contained the same number of levels. Crucially, no variable encoded quantitative trade-offs (e.g., probabilities), ensuring that differences across domains were not driven by numerical computation. Although kidney allocation scenarios appeared longer due to a constant reference patient, the varying information was matched in length and structure across domains. All variables were non-numerical and presented in paragraph form.

After completing 15 questions in their assigned condition, participants rated the extent to which they relied on a specific decision process (1 = "Not at all" to 7 "Completely"). Participants then provided a written explanation of their decision-making process, which would be shown to another participant

tasked with predicting their responses. They were instructed to write explanations that would enable accurate prediction and were incentivized with a performance-based bonus. There were no significant differences between conditions in description length ($M = 63$ - 64 words, $t(187.58) = 0.09$, $p = .925$) or majority class percentage ($M = 64\%$ in both, $t(27.79) = -0.24$, $p = .812$). Full stimulus details are available in the repository.

Matching Phase

In the Matching Phase, a new set of participants received materials from the Decision Phase, with the condition determining whether they saw only the original stimuli (Uninformed) or both the stimuli and the decision-maker's explanation (Informed).

Informed participants were presented with the same moral scenarios as their partner, accompanied by a description of how their partner explained their decision-making process for those scenarios. They were then asked to use this description to try to predict the other participant's decisions.

Uninformed participants were also paired with a Decision Phase participant (though they were not told this) and presented with the same moral scenarios. However, they received no mention of a partner or any decision-making process. Instead, they made their own independent decisions on the scenarios. Their responses serve as the baseline agreement (i.e., the expected overlap between two randomly paired participants) level between any two randomly paired participants—a point of comparison for the Informed condition.

Participants We recruited a total of 404 adult participants from Prolific to complete an online survey through Qualtrics. Of these, 27 responses were excluded due to a failed comprehension check which resulted in a final sample of 377 participants ($M_{age} = 39.9$, $SD_{age} = 13.5$; female = 207, male = 166, other = 4; White = 68.2%, Black or African American = 9.3%, Multiracial/Other = 11.9%, Asian or Asian American = 7.7%, Hispanic or Latine = 2.9%).

Informed Condition Participants in the Informed condition were told they had been matched with a prior participant and were shown an example item. They then received their partner's explanation of their decision process and were instructed to use it to predict the partner's responses across 15 decisions, with a performance-based bonus for accuracy. Below is an example question:

***Patient A:** Jesse is a part-time librarian who generally follows doctor's advice but occasionally overlooks minor recommendations. She relies on bi-weekly treatments to manage kidney function and prevent decline. She feels a quiet determination to improve her health, acknowledging past missteps but focusing on moving forward. She has a history of mostly adhering to medical advice, with minor lapses during stressful times.*

***Patient B:** Charlie is a community volunteer. He shows little concern for maintaining a healthy lifestyle, often neglecting his well-being entirely. He has a history of*

ignoring medical advice regarding his condition but has recently shown efforts to improve health management. He has developed multiple organ complications due to kidney failure and is at risk of permanent disability. He accepts his condition, focusing on making the most of his current situation.

Their decision process: When deciding on who should receive the kidney transplant I would consider: (1) Which patient has the worst health at the moment. (2) Which patient shows more concern for their own health and takes responsibility for their own health. (3) Which patient is more negligent towards taking medical advice. (4) Which patient is more remorseful about their own past choices and how they had an impact on their current health issues.

Based on the other participant's decision-making process, which patient should receive the kidney?

Uninformed Condition Participants in the Uninformed condition read the same scenarios as a paired Decision Phase participant but received no information about a partner or decision process and made independent decisions. Their responses provide a baseline measure of agreement for comparison with the Informed condition.

Results

We report three main findings: (1) participants reported greater reliance on a specific decision process in the kidney allocation domain than in the forgiveness domain, (2) baseline agreement was significantly higher in the kidney allocation domain than in the forgiveness domain, and (3) explanations increased matching accuracy in both domains, with a larger effect in forgiveness, yielding similar overall matching rates when explanations were provided.

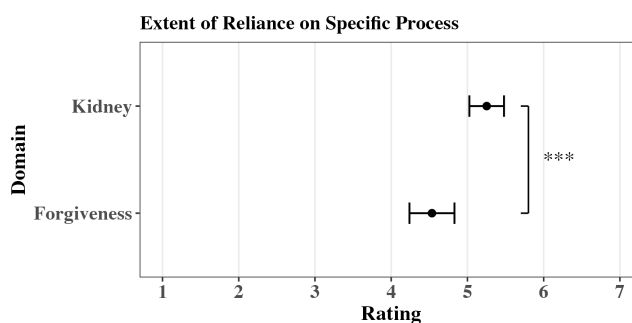


Figure 2: Rating of extent to which participants relied on a specific process when making decisions across domains. Points indicate means and vertical lines indicate 95% confidence intervals.

Participants report greater extent of reliance on a specific process in the kidney allocation condition. Participants' self-reported reliance on a specific process when making decisions was significantly higher in the kidney allocation condi-

tion ($M = 5.3$, $SD = 1.1$) compared to the forgiveness condition ($M = 4.5$, $SD = 1.5$), ($t(184.6) = -3.78$, $p < 0.001$, 95% CI for the mean difference $[-1.1, -0.3]$, Cohen's $d = 0.54$). Results are shown in Figure 2. This result suggests that kidney allocation decisions were experienced as more structured and rule-based than forgiveness decisions.

Baseline level of agreement is significantly higher for the kidney allocation condition. We used a generalized linear mixed-effects model with a binomial logit link function to analyze the number of matched answers with respect to moral domain, Informed versus Uninformed condition, and trial number. Results are shown in Figure 3. The model included domain (kidney allocation or forgiveness), condition (informed or uninformed), trial number, and their interactions as fixed effects, and participants as a random intercept.

There was a main effect of domain, $\chi^2(1) = 5.05$, $p = .025$, condition, $\chi^2(1) = 68.89$, $p < .001$, and their interaction, $\chi^2(1) = 6.71$, $p = .010$, but no main effect of trial number. Decomposing the interaction with estimated marginal means showed that participants in the Uninformed Condition differed significantly in matched answers across the kidney allocation and forgiveness domains ($\Delta = 1.723 \pm 0.35$, $p < .001$, $d = 0.54$), suggesting that different moral domains have different levels baseline agreement between any given pair of participants.

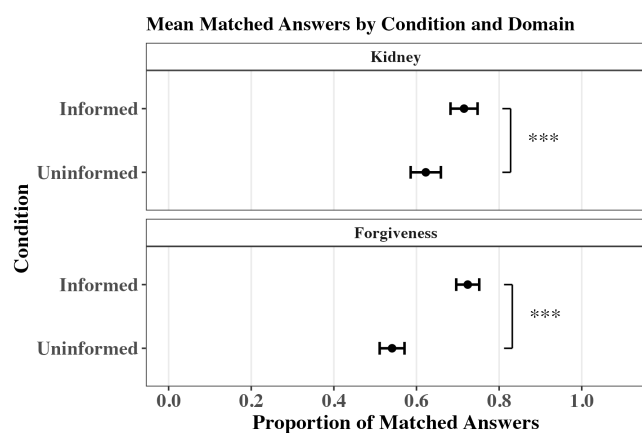


Figure 3: Proportion of matched answers across conditions. Points indicate means and vertical lines indicate 95% confidence intervals.

Explanations had a larger effect in the forgiveness condition, resulting in similar matching accuracy across domains in the Informed Condition. Using the same model and estimated marginal means, we found that there was an increase in matched answers from the Uninformed Condition to the Informed Condition for both the kidney allocation ($\Delta = 1.38 \pm 0.34$, $p < .001$, $d = 0.61$) and forgiveness ($\Delta = 2.76 \pm 0.34$, $p < .001$, $d = 1.22$) domains, and that the increase was larger for the forgiveness domain. However, the number of matched answers across domains in the Informed Condition was not different. These results suggest that the effectiveness

of the decision-making descriptions in increasing the number of matched answers depends on the moral domain, but matching accuracy converged across domains when explanations were provided.

Exploratory Analysis: Perceived reliance on specific rules did not predict explanation effectiveness. In an exploratory analysis, we examined whether participants' self-reported reliance on specific rules correlated with matching accuracy in the Informed condition. There was no significant correlation overall ($r = -.14$, $p = .056$). However, examining domains separately revealed a negative correlation in the kidney allocation domain ($r = -.31$, $p = .003$) and no correlation in the forgiveness domain ($r = .04$, $p = .728$). The absence of a positive relationship—and particularly the unexpected negative correlation in kidney allocation—suggests that participants could not accurately judge how effectively their reasoning could be communicated to others.

Discussion

This study investigated whether normative expectations for justification across moral domains influence how effectively people can articulate their decision-making processes in ways that others can use. We compared two contrasting moral domains that differ in their normative expectations for justification. Participants made a series of moral decisions in one of the two domains, rated their reliance on a specific decision process, and provided written explanations of their reasoning. These explanations were then given to new participants who attempted to use them to match the original decisions, or to participants who made decisions independently without any explanation.

Key Findings and Theoretical Implications

First, participants perceived themselves as relying more on a specific decision process in the kidney allocation domain than in the forgiveness domain, though the effect size was moderate and the absolute difference was relatively small. Both groups reported moderate-to-high reliance on specific processes, suggesting that participants across both domains felt they were following some kind of systematic approach.

Second, baseline agreement was significantly higher in the kidney allocation domain than in the forgiveness domain. This likely reflects the fact that kidney allocation is governed by more widely shared moral principles that constrain individual variation, while forgiveness decisions are more idiosyncratic and shaped by personal values.

Third, explanations increased matching accuracy in both domains relative to baseline, but the increase was larger in the forgiveness domain, where baseline agreement was lower. Most surprisingly, despite these different pathways—with forgiveness showing a larger gain from explanations—matching accuracy in the Informed Condition was similar across both domains, with participants achieving approximately 11 out of 15 matches in both conditions.

Finally, an exploratory analysis showed that perceived reliance on specific rules did not predict how well explanations enabled others to reproduce participants' decisions, suggesting that participants' metacognitive judgments about how rule-based their decisions were did not reliably track how informative their explanations actually were.

This pattern invites several interpretations. One possibility is that explanations in both domains successfully captured decision-relevant information, but through different mechanisms. In kidney allocation, explanations may have articulated principles that were already partially shared, making them easier to apply. In forgiveness, explanations may have introduced novel frameworks that reduced individual variability by imposing structure where little existed before. Both pathways could result in similar matching accuracy—one by reinforcing existing norms, the other by creating new ones.

Alternatively, people may possess a general capacity to generate coherent, usable explanations when prompted, regardless of domain or underlying reasoning structure. This interpretation is supported by the null correlation between perceived rule use and explanation effectiveness: participants may construct post hoc justifications that are useful even when they do not correspond to the actual basis of their judgments, aligning with the Social Intuitionist Model's (Haidt, 2001) emphasis on post hoc reasoning. This finding is notable given evidence that people confidently generate explanations that do not reflect their actual decision processes (Johansson, 2005), yet our results suggest such explanations can nonetheless be pragmatically useful. Cushman (2020) extends this view, suggesting that rationalization may itself be rational and adaptive, allowing individuals to extract useful principles from intuitive processes. Participants may have been constructing coherent decision rules to interpret and communicate their actions, and these constructions were effective in guiding others' predictions, even if they did not accurately describe the original processes. Fine (2006) offers a complementary perspective, proposing that automatic moral judgments may reflect internalized prior reasoning. Together with the null correlation between perceived rule use and explanation effectiveness, these findings indicate that explanations can be pragmatically useful even in the absence of accurate introspective access.

Regarding domain differences: baseline agreement varied substantially, and explanations had larger effects in forgiveness. This does not necessarily indicate that forgiveness decisions lacked internal structure, but rather that individual decision-makers relied on more idiosyncratic principles not widely shared across the population. Explanations may have provided a common framework that aligned diverse individual structures. Despite these different pathways—kidney allocation explanations reinforcing existing norms, forgiveness explanations creating shared structure from diverse individual approaches—both produced similar final matching accuracy. This convergence implies explainability depends less on domain structure than on general capacity to construct coherent, communicable accounts when prompted.

Limitations and Future Directions

While our study introduces a novel measure of moral explainability, several limitations warrant consideration. First, participation in a psychology study may itself induce expectations that one's reasoning should be explainable, even in the absence of strong normative pressure from the task. Anticipation of evaluation may have encouraged more structured reasoning, potentially attenuating differences between high- and low-justification domains. Importantly, participants were not informed of the explanation requirement until after completing all decisions, which limits the degree to which anticipation of justification could have shaped decision behavior itself. However, it remains possible that domain-level expectations—participants' prior beliefs about whether kidney allocation or forgiveness decisions should be principled—influenced how decisions were experienced and subsequently explained. Future work could address this by manipulating the explicitness of explanation demands or by comparing contexts where participants do and do not expect their reasoning to be communicated. Second, our study was restricted to two moral domains. Although chosen to contrast justification norms, differences in their stimuli may have introduced confounds, and the findings may not generalize to other moral contexts. Future research could test whether similar patterns emerge in additional domains with varying expectations for justification, such as criminal sentencing, legal pardons, dietary choices, or loyalty dilemmas.

Conclusion

This study investigated whether normative expectations for justification across moral domains affect how well people can articulate their decision-making processes in ways that others can use. We found that although domains differ in baseline agreement and perceived reliance on rules, explanations in both high- and low-justification domains ultimately enabled similar levels of decision prediction. This suggests that moral explainability may depend less on the inherent structure of a domain and more on individuals' general capacity to construct coherent, communicable accounts of their reasoning. These findings have important implications for theories of moral cognition: If moral explanations are constructed in response to the demand for justification rather than retrieved from internal processes, then we must critically examine the assumption that justifications are transparent reflections of decision processes. Instead, we should focus on understanding when, how, and why explanations are constructed, and on evaluating their functional effectiveness in enabling communication and decision reproduction. Understanding moral explainability as a communicative and coordinative act, rather than as a window into cognitive processes, offers a more realistic foundation for both theoretical advances in moral psychology and practical applications in contexts requiring moral accountability and transparency.

Acknowledgments

We thank Joshua Knobe and Tom McCoy for their invaluable mentorship and feedback throughout the early stages of this project. This work was supported by the Yale Cognitive Science Program.

References

- Andrews, M. (2000). Forgiveness in Context. *Journal of Moral Education*, 29(1), 75–86. <https://doi.org/10.1080/030572400102943>
- Cowley, Christopher. (2010). Why Genuine Forgiveness must be Elective and Unconditional. *Ethical Perspectives*, (4), 556–579. <https://doi.org/10.2143/EP.17.4.2059846>
- Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, 43, e28. <https://doi.org/10.1017/S0140525X19001730>
- Cushman, F., Young, L., & Hauser, M. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm. *Psychological Science*, 17(12), 1082–1089. <https://doi.org/10.1111/j.1467-9280.2006.01834.x>
- De Fijter, J. W. (2010). Kidney allocation: Where utility and fairness meet. *Nephrology Dialysis Transplantation*, 25(6), 1746–1749. <https://doi.org/10.1093/ndt/gfq164>
- FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, 123(3), 434–441. <https://doi.org/10.1016/j.cognition.2012.02.001>
- Fine, C. (2006). Is the emotional dog wagging its rational tail, or chasing it?: Reason in moral judgment. *Philosophical Explorations*, 9(1), 83–98. <https://doi.org/10.1080/13869790500492680>
- Gismero-González, E., Jódar, R., Martínez, M. P., Carrasco, M. J., Cagigal, V., & Prieto-Ursúa, M. (2020). Interpersonal Offenses and Psychological Well-Being: The Mediating Role of Forgiveness. *Journal of Happiness Studies*, 21(1), 75–94. <https://doi.org/10.1007/s10902-018-00070-x>
- Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), 322–323. <https://doi.org/10.1016/j.tics.2007.06.004>
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 293(5537), 2105–2108. <https://doi.org/10.1126/science.1062872>
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834. <https://doi.org/10.1037/0033-295X.108.4.814>
- Haidt, J. (2003). The moral emotions. In *Handbook of affective sciences* (pp. 852–870). Oxford University Press.
- Hannikainen, I. R., Tobia, K. P., De Almeida, G. D. F. C. F., Struchiner, N., Kneer, M., Bystranowski, P., Dranseika, V., Strohmaier, N., Bensinger, S., Dolinina, K., Janik, B., Lauraitytė, E., Laakasuo, M., Liefgreen, A., Neiders, I., Próchnicki, M., Rosas, A., Sundvall, J., & Żuradzki, T. (2022). Coordination and expertise foster legal textualism. *Proceedings of the National Academy of Sciences*, 119(44), e2206531119. <https://doi.org/10.1073/pnas.2206531119>
- Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science*, 310(5745), 116–119. <https://doi.org/10.1126/science.1111709>
- Kahn, B. E., & Baron, J. (1995). An Exploratory Study of Choice Rules Favored for High-Stakes Decisions. *Journal of Consumer Psychology*, 4(4), 305–328. https://doi.org/10.1207/s15327663jcp0404_01
- Kunreuther, H., Meyer, R., Zeckhauser, R., Slovic, P., Schwartz, B., Schade, C., Luce, M. F., Lippman, S., Krantz, D., Kahn, B., & Hogarth, R. (2002). High Stakes Decision Making: Normative, Descriptive and Prescriptive Considerations. *Marketing Letters*, 13(3), 259–268. <https://doi.org/10.1023/A:1020287225409>
- McCullough, M. E. (2001). Forgiveness: Who Does It and How Do They Do It? *Current Directions in Psychological Science*, 10(6), 194–197. <https://doi.org/10.1111/1467-8721.00147>
- Moll, J., De Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The Neural Correlates of Moral Sensitivity: A Functional Magnetic Resonance Imaging Investigation of Basic and Moral Emotions. *The Journal of Neuroscience*, 22(7), 2730–2736. <https://doi.org/10.1523/JNEUROSCI.22-07-02730.2002>
- Morris, A., Carlson, R. W., Kober, H., & Crockett, M. (2023, October). Introspective access to value-based multi-attribute choice processes. <https://doi.org/10.31234/osf.io/2zrfa>
- Paxton, J. M., & Greene, J. D. (2010). Moral Reasoning: Hints and Allegations. *Topics in Cognitive Science*, 2(3), 511–527. <https://doi.org/10.1111/j.1756-8765.2010.01096.x>
- Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and Reasoning in Moral Judgment. *Cognitive Science*, 36(1), 163–177. <https://doi.org/10.1111/j.1551-6709.2011.01210.x>
- Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review*, 110(1), 193–196. <https://doi.org/10.1037/0033-295X.110.1.193>
- Worthington, E. L., Witvliet, C. V. O., Pietrini, P., & Miller, A. J. (2007). Forgiveness, Health, and Well-Being: A Review of Evidence for Emotional Versus Decisional Forgiveness, Dispositional Forgiveness, and Reduced Unforgiveness. *Journal of Behavioral Medicine*, 30(4), 291–302. <https://doi.org/10.1007/s10865-007-9105-8>