What do moral rules mean?

Joe Kwon[1], Lio Wong[1], Josh Tenenbaum[1], Sydney Levine (sydney.levine@nyu.edu)[2]
[1]Department of Brain and Cognitive Sciences, MIT
[2]Allen Institute for AI

Abstract

People often communicate social and moral expectations to one another in terms of rules. While these rules often seem simple (e.g. "No walking on the grass") people understand that much more is being communicated than it at first appears. In this paper we present a novel theory of how people understand moral rules and instantiate that theory in a computational cognitive model. We argue that moral rules are a special kind of speech act, directed at a group and expressing a summary of the agreement that rational actors would arrive at when navigating an interdependent choice problem. Rules, on this conception, are therefore closely tied to their *reasons*, which reflect the interests of the parties to the agreement. This structure is very powerful: it allows people to interpret and apply rules flexibly in novel situations and edge cases. Specifically, we argue that when judging whether an action violates a rule, people reflect on the reason for the rule and harness mental models of agreement (and their heuristic approximations) to determine whether the action upholds or undermines the reason. Put simply, people ask whether the purpose of the rule would still be upheld if the rule permitted everyone to act in this way.

Keywords: Computational Modeling; Moral cognition

Introduction

Rules permeate our physical and virtual environments, from campus access policies to medical protocols to simple park directives. While rules may appear straightforward, they often convey complex information beyond their literal meaning. Consider a park sign stating "Please keep off the grass!" This simple directive implicitly permits certain exceptions (like gardeners) while prohibiting unstated activities (like pogo-sticking), revealing how rules communicate both explicit restrictions and implicit understanding about the world they govern.

This inherent complexity in rule interpretation also presents a central challenge in AI safety: how do we effectively communicate human social and moral rules to AI systems? Without a computational framework for human rule understanding, current approaches to AI alignment remain indirect and potentially unreliable. Recent work has highlighted the need for robust, algorithmic specifications of rules (Zhi-Xuan, Carroll, Franklin, & Ashton, 2024; Klingefjord, Lowe, & Edelman, 2024; Dalrymple et al., 2024), yet such characterizations remain elusive. Our approach posits that rules exist for specific reasons and function as approximate implementations of their underlying purposes. By modeling both a rule's purpose and the generative mechanism that transforms complex intentions into simple directives, we can better understand how humans interpret and apply rules.

What do moral rules mean? Linguistic, legal, and social communicative approaches

Traditional approaches view rules as linguistic imperatives that constrain policies or define utilities (Portner, 2007). However, community-directed rules that promote collective goals (such as laws in a democracy or formal or informal policies guiding an institution) reveal additional complexity. Legal theory traditions have long grappled with the myriad of ways to understand such directives. We take inspiration from the "fuctionalist" school of thought which posits that meaningful rule interpretation requires understanding a rule's underlying *purpose* (Hart, 1958; Fuller, 1958; Schauer, 1987). We build on this theoretical precedent as well as cognitive frameworks that model communicative inferences through intuitive Rational Speech Act - social theories (Frank & Goodman, 2012; Goodman & Frank, 2016; Degen, 2023). While previous work has focused on dyadic communication, we extend these approaches to community-directed rules.

Connecting rules to their reasons

Recent cognitive science research suggests that moral judgment mechanisms are approximations of morality's fundamental function: finding mutually beneficial solutions to multi-agent decision problems (Levine, Chater, Tenenbaum, & Cushman, 2024; André, Debove, Fitouchi, & Baumard, 2022). This "resource-rational contractualism" framework proposes that while direct negotiation would be ideal, people employ various approximations based on available cognitive resources (Lieder & Griffiths, 2020; Wu, Ren, Gerstenberg, Choi, & Levine, 2024; Trujillo, Zhang, Zhi-Xuan, Tenenbaum, & Levine, 2024). Within this framework, rules can be understood as simplified outputs of more resource-intensive bargaining processes guided by a particular objective: to advance the interests of the parties to the bargain. Rule interpretation therefore can involve reconstructing the generative process in order to handle a potential edge case. For instance, this can work by simulating what would happen if everyone felt at liberty to completely violate the rule in this case and whether such a rule emendation would be agreed upon (Levine, Kleiman-Weiner, Schulz, Tenenbaum, & Cushman, 2020; Kwon, Tan, Tenenbaum, & Levine, 2023). Alternately, someone could simulate the particular policies that

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

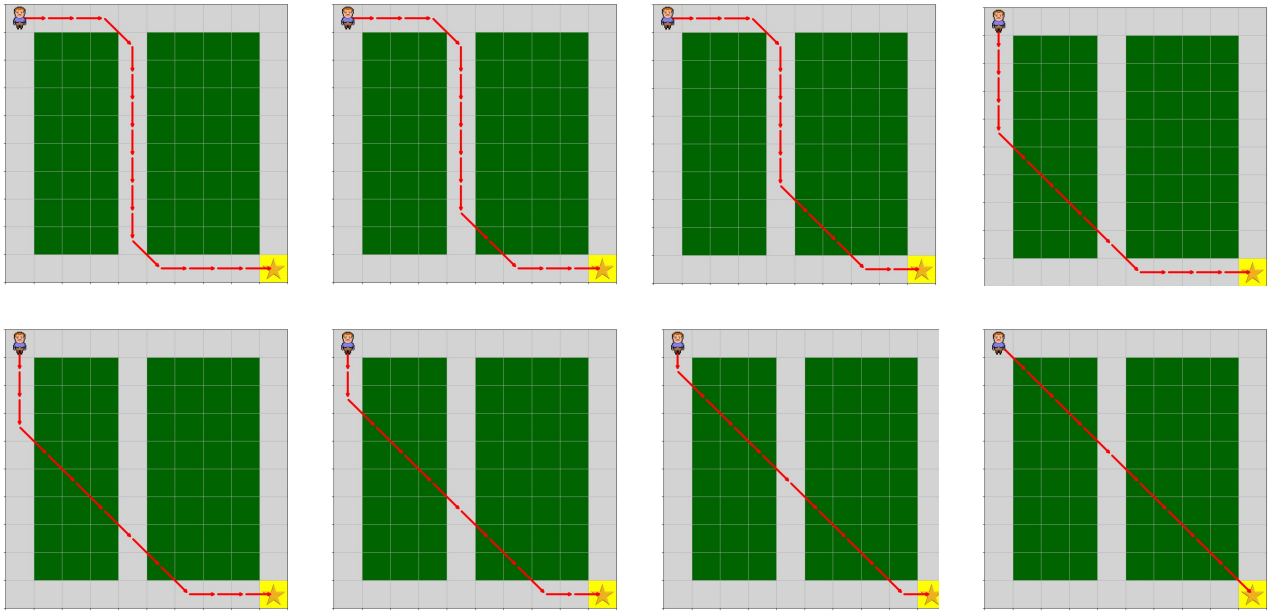

Figure 1: Varying trajectories taken by an agent towards a goal (yellow star), ranging from *no shortcut* and walking completely on the sidewalk (top left) to *maximally shortcutting* over the grass from their initial position to the goal location.

would be ideal for each agent involved, guaranteeing agreement (Levine, Kleiman-Weiner, Chater, Cushman, & Tenenbaum, 2024; Misyak & Chater, 2014). The present approach explores a middle ground: asking what would happen if all agents in a particular circumstance broke the rule *to a certain extent*.

Computational Framework

We now present a computational framework that formalizes how people figure out what the purpose of a rule is and then use that purpose to reason about exceptions to the rule.

Rules and collective action problems in multi-agent environments We focus on publicly communicated rules directed at communities, considering a setting with n individual agents $A_i \in \{A_0 ... A_n\}$ acting in a shared world W.

Individual agents make decisions based on two utility components:

- Short-term, individual utility $U_{A_i,immediate}$ specific to each agent A_i, capturing immediate goals like navigation with effort-based costs
- Long-term utility $U_{A_i,background}$ reflecting general background utilities contingent on the shared world W, such as environmental quality

An agent's overall utility is defined as $U_{overall} = \lambda U_{A_i,immediate}\ 1 - \lambda U_{A_i,background}$, where λ captures the balance between immediate goals and longer-term considerations. Agents form plans $\Pi_{A_i} = \{\alpha_{A_i,t_0} ... \alpha_{A_i,t_j}\}$, with world dynamics collectively impacted by all agents' actions through a transition function $W_{t_j} \rightarrow \langle \alpha_{A_i,t_j} \in \{A_0 ... A_n\} \rangle \rightarrow W_{t_{j1}}$.

Collective action problems arise when agents cannot independently optimize their utilities without impeding others. While explicit and virtual bargaining solutions (Levine, Kleiman-Weiner, et al., 2024; Trujillo et al., 2024; Misyak & Chater, 2014) attempt to solve for optimal joint plans $\Pi_{A_i} \in \{A_0 ... A_n\}$, this approach faces practical challenges in coordinating multiple agents' actions.

Interpreting and making exceptions to rules based on universalization Agents judge whether to break a rule by considering if their immediate need crosses a threshold, then evaluating the effects if all community members applied the same threshold for exceptions. This approach balances strict rule-following with necessary exceptions, capturing intuitions about when rules should be broken for greater mutual benefit.

Formally, an agent Π_{A_i} chooses between a strict rule-following policy $\Pi_{A_i,rule}$ and an exception policy $\Pi^*_{A_i,exception}$. Given their immediate utility $U_{A_i,immediate}$, they reason about a universalized exception-granting policy where agents follow $\Pi_{A_j,exception}$ `iff` $U_{A_j,immediate} \geq U_{A_i,immediate}$ `else` $\Pi_{A_i,rule}$. Using their prior distribution $PU_{A_i,immediate}$ on other agents' utilities, they can estimate the cumulative utility of universalizing this approach.

Inferring the purpose of rules based on communicative informativity Agents can reason about the environment W assuming rules communicate policies that improve mutual benefit over individual planning. Let $\Sigma_{A_i} EU_{A_i} | \Pi_{A_i}, W$ represent the expected cumulative mutual utility given world W and policies Π_{A_i}. Rules' utterance utility U_{rule} increases with their ability to improve expected mutual benefit through rule interpretation (including exceptions) compared to myopic planning.

Following speaker models from (Goodman & Frank, 2016), rule production can be modeled as approximately maximizing utterance utility. Since this generative model depends on W, observing a particular rule provides evidence for inferring components of W. This captures the intuition that speakers are less likely to communicate rules that yield no improvement over myopic action.

Our test case: *keep off the grass* in a community park

We illustrate our framework through a concrete example: a *keep off the grass* rule posted in a community park. Our model environment is a grid-world with sidewalks, grass, and $k = 6$ distinct perimeter destinations: an ice cream truck, police car, temporary vaccination clinic, porta-potty, passing friend, and a person screaming in pain. These destinations vary in both their immediate utility and the frequency with which community members might need to reach them. The key behavioral variable is the extent to which agents cut across the grass to reach their destinations.

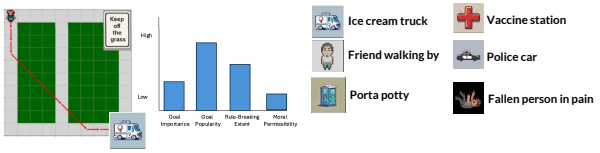

Figure 2: (Left) Example grid-world scenario where an agent partially crosses grass to reach an ice cream truck, despite posted rules. (Center) Key variables affecting moral judgments: goal importance and inferred goal frequency inform permissibility of rule-breaking. (Right) Various destinations eliciting different intuitions about goal importance and expected community usage patterns.

Our model makes two key predictions: (1) judgments about rule-breaking acceptability depend on an agent's behavior relative to community needs, and (2) the rule's existence implies information about the environment—specifically, grass fragility. We formalize this through agents who reason about rule permissibility by asking: *What if everyone in my situation (or one more urgent) felt at liberty to break the rule to this extent?* The following sections detail how we transform this question into a computational model incorporating community goals, behaviors, and outcomes.

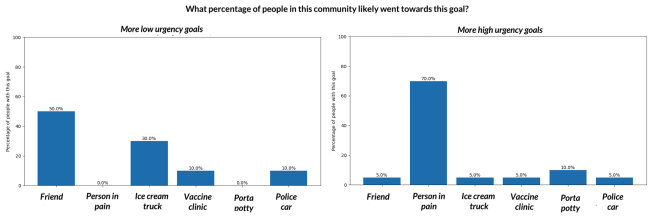

Figure 3: Specific communities with different simulated distributions over the number of individuals in the community who tend to go to varying destinations. In the *more low urgency* goals condition (left), relatively more of the community tends to pursue lower urgency goals on a given day than the normalized distribution over goal frequencies elicited in our baseline experiment. In the *more high urgency* goals condition (right), relatively more of the community tends to pursue higher urgency goals.

Computational Model for Walking on Grass

We developed a computational model simulating agents navigating a grid world under a rule prohibiting walking on grass, examining how agents balance individual goals against collective consequences. Our model incorporates a myopia parameter representing how much agents prioritize immediate benefits over long-term community effects, allowing us to explore how decisions align with universalization principles.

Full Generative Model

Agents and Goals Agents start from random sidewalk locations and pursue randomly assigned goals, with selection probabilities based on empirical data about frequency of different goals in the community. Each goal has a utility value U_{goal} sampled from a normal distribution, with means and standard errors derived from human judgments (e.g., $U_{\text{friend}} = 52.7 \pm 3.87$).

Planning and Trajectory Generation Agents employ two strategies:

Rule-Following Trajectory: The shortest path adhering to sidewalks, computed using Breadth-First Search (BFS).

Shortcut Trajectory: Paths that may cross grass, parameterized by a *degree of rule-breaking* ($\sigma \in 0, 1$), where $\sigma = 0$ indicates full compliance and $\sigma = 1$ represents complete disregard for the rule.

Utility Calculation Agent utility is computed as:

$$U_{\text{agent}} = U_{\text{goal}} - c \cdot L_{\text{path}}, \tag{1}$$

where U_{goal} is the goal utility, c is the path cost (set to 1), and L_{path} is the path length, with diagonal movements costing $\sqrt{2}$ units.

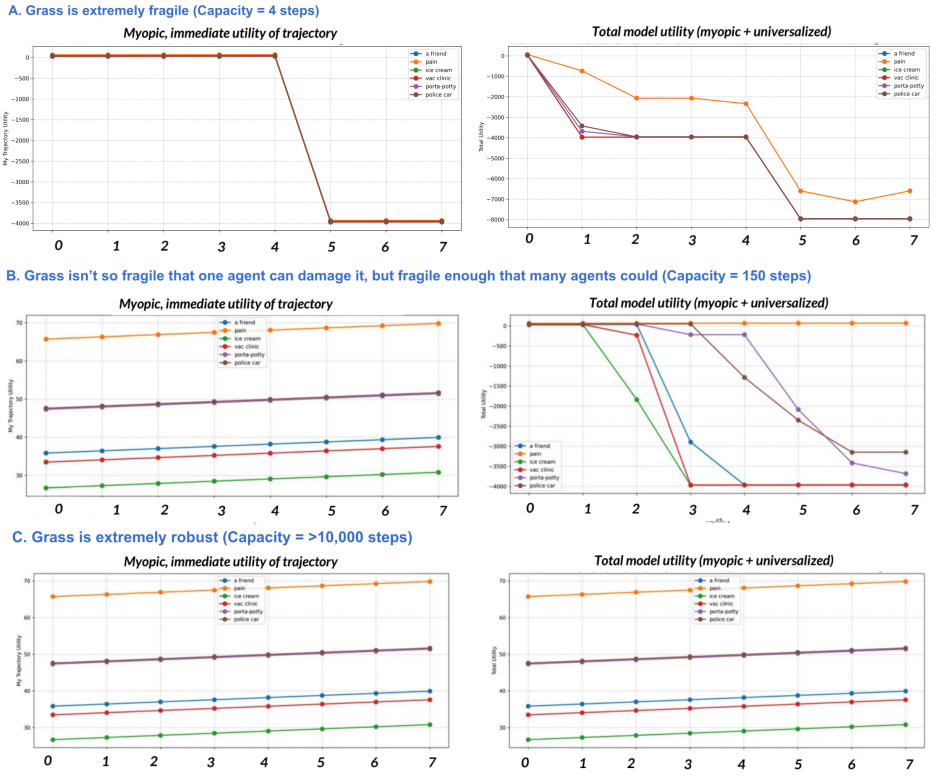

Figure 4: Comparative *myopic, short-term observable* and *total (myopic and universalized)* utilities over agent trajectories, given different hypothetical values for the *fragility of the grass* – its capacity to withstand a number of total cumulative number of steps in any given day before incurring costly damage. (A, top) If the grass is *extremely fragile*, such that any one agent could immediately observe costly damage directly from their actions, the publicly posted sign serves little value relative in the context of communicating relevant directives for the community that agents could not observe on their own. (B, middle) If the grass is not so fragile that it could be damaged by any one agent, but could incur collective damage, the sign is *highly informative* about how many agents might consider revising their actions beyond what they could immediately observe. (C, bottom) If the grass is *extremely robust*, the publicly posted sign again serves little value in communicating any important community cost.

Universalization and Community Cost Agents simulate community impact by considering a population of size N, where simulated agents have goal utilities equal to or greater than the universalizing agent's. The grass has capacity C_{grass} before destruction, incurring cost $U_{\text{grass\_cost}}$.

The universalized utility is:

$$U_{\text{univ}} = \frac{1}{N}\left(\sum_{i=1}^{N} U_{\text{agent},i} - U_{\text{grass\_cost}} \cdot \mathbb{I}\left(S_{\text{total}} > C_{\text{grass}}\right)\right), \quad (2)$$

where S_{total} is total grass steps taken and $\mathbb{I}$ is the indicator function.

Myopia Parameter Agents balance personal and universal utility through myopia parameter $\lambda \in 0, 1$. The overall utility is:

$$U_{\text{overall}} = \lambda U_{\text{agent}}\ 1 - \lambda U_{\text{univ}}. \quad (3)$$

A fully myopic agent ($\lambda = 1$) maximizes personal utility, while a fully farsighted agent ($\lambda = 0$) prioritizes collective outcomes.

Inference of the Myopia Parameter To predict permissibility judgments, we infer the myopia parameter λ that would make the observed behavior optimal. Given a trajectory with a rule-breaking degree σ_{obs}, we compute the posterior probability:

$$P\lambda|\sigma_{\text{obs}} = \frac{\exp\left(U_{\text{overall}}\lambda, \sigma_{\text{obs}}\right)}{\lambda' \exp\left(U_{\text{overall}}\lambda', \sigma_{\text{obs}}\right)}. \quad (4)$$

This framework connects observed behavior with moral evaluation: Lower inferred λ values indicate more community-oriented agents, while higher values suggest more self-serving behavior judged less permissible.

Model simulations and results

We present results from simulations examining how agents navigate towards different goals with varying degrees of grass shortcuts. These simulations demonstrate how moral judgments about rule exceptions arise from agent behavior in community contexts, and how rules communicate information about the environment itself.

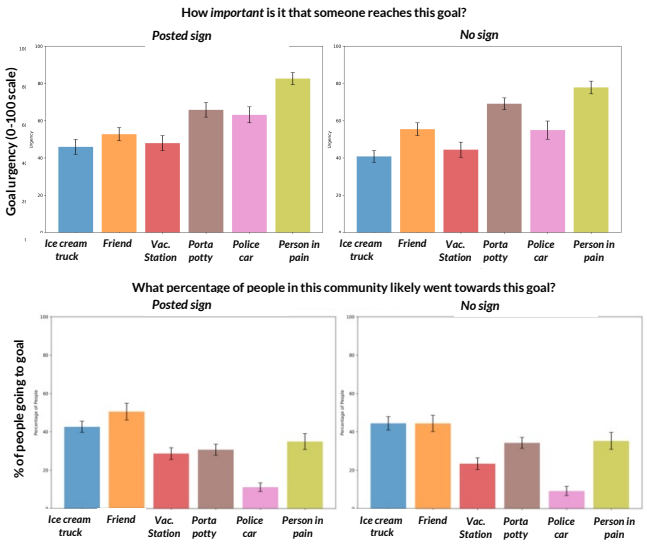

Figure 5: Human elicited background priors about (top) the relative *urgency* of particular destinations for any given agent in the community, and the *expected percentage of people* going towards any given destination out of the community.) Note that these percentages do not sum to one as participants were told that agents could pursue multiple goals on a given day.

Simulation: Walking on the Grass Towards Various Goals With and Without a Posted Rule

Our first simulation compares conditions with and without an explicit "Do not walk on the grass" rule. Agents navigate toward six destinations using eight distinct trajectories (Fig. 1), with shortcutting quantified by percentage of path length saved.

Human Priors About Goal Frequency and Urgency To parameterize the model, we collected judgments from 50 Prolific participants about goal importance and frequency. Participants estimated both the percentage of community members pursuing each goal and each goal's relative urgency. As shown in Fig. 5, goals varied significantly in both urgency and frequency, with these judgments remaining consistent regardless of whether the rule was posted. This invariance suggests that the rule itself doesn't alter baseline expectations about community behavior, but rather influences how these behaviors are interpreted.

Moral Judgments and Permissible Exceptions Our results reveal three key patterns in moral judgments (Fig. 6). First, even minimal rule violations suggest significant myopia, as shown by sharp increases in inferred myopia for small deviations. Second, this effect scales with goal urgency—cutting across grass for an ice cream truck shows higher myopia than identical trajectories toward emergency situations. Third,

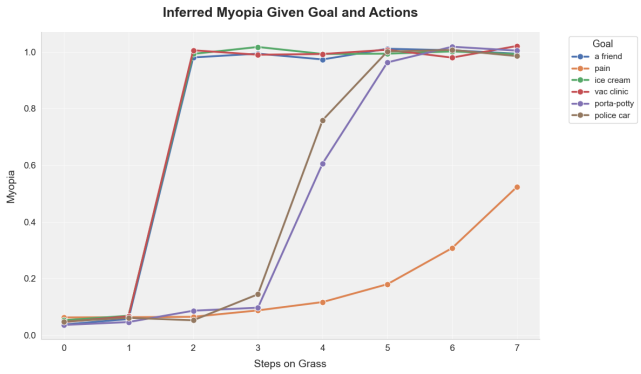

Figure 6: Inferred *myopia* coefficient over an agent utilities – how much an agent takes into account short-term, immediate utilities like the utility of their specific goal and desire to get their as efficiently as possible, as opposed to a *universalized, long-term* coefficient indicating their weighting for long-term community costs by imagining everyone else in the community affording themselves of the same right and degree to a shortcut given the urgency of their goals. These two coefficients are normalized to sum to 1, so coefficients indicate relative weight. Intuitively, inferred myopias are *inversely* related to the perceived permissibility of a given trajectory (amount of steps on the grass) given the rule and relative frequency and urgency with which agents in the community at large pursue their own goals.

community context strongly modulates these judgments: actions appear more myopic when many community members have equally or more urgent goals.

Inferences About Environmental Properties Our analysis of grass fragility reveals that rules communicate crucial information about environmental properties (Fig. 4). The data show that rules are most informative in intermediate fragility conditions—where collective but not individual actions cause damage. With extremely fragile grass (damage from single steps) or highly robust grass (no damage from many steps), the rule provides little information beyond what agents could observe directly. This finding demonstrates how rules implicitly communicate environmental properties through their function in coordinating collective behavior.

Keep off the grass in communities with different distributions of goals

Our second simulation demonstrates how identical actions can carry different moral weight depending on community composition (Fig. 3). We compared two conditions: communities with predominantly low-urgency versus high-urgency goals.

The results (Fig. 7) reveal a striking context effect: identical actions suggest different degrees of myopia depending on community needs. When an agent cuts across grass to reach a porta-potty, this behavior appears significantly less myopic

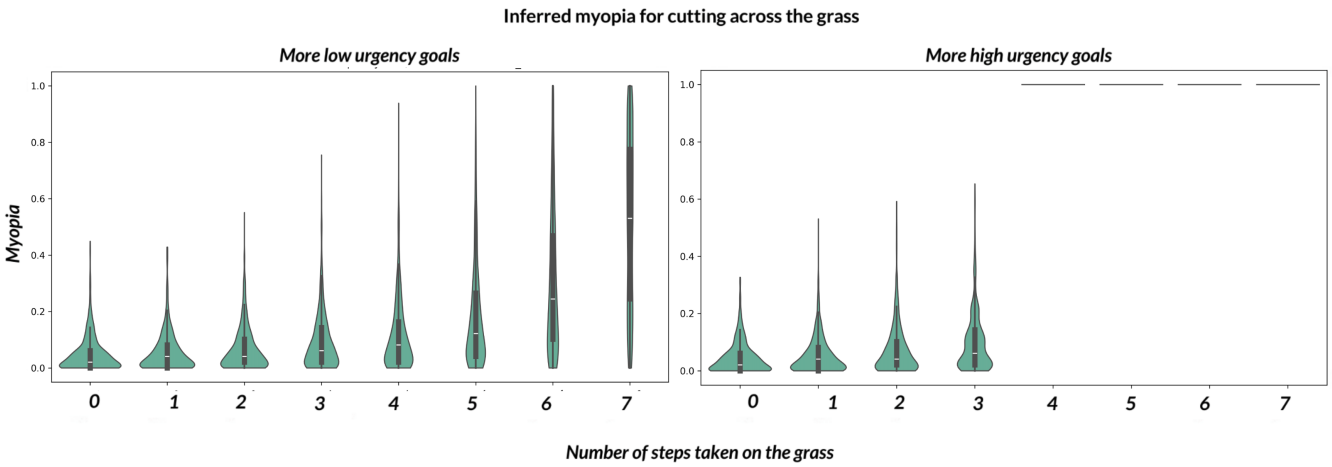

Figure 7: Inferred *myopia* coefficient over the same agent trajectories in different communities. All trajectories consider an agent whose destination is the *porta-potty*. (Left) In the community where other people tend to go to less urgent goals in general, cutting across the grass towards the porta potty connotes *less myopia* in general relative to the same trajectories in the other condition. (Right) In the community where other people tend to go to *more* urgent goals relative to the community as a whole, cutting across the grass connotes *more* myopia.

in communities with predominantly low-urgency goals. However, in communities where more people pursue high-urgency goals, the same shortcut indicates greater myopia, as it suggests less consideration for others with similar or more pressing needs.

Discussion

Our computational framework demonstrates how humans may interpret rules through the lens of their underlying purposes and functions in promoting collective benefit. The model builds on recent work in cognitive science suggesting that moral judgment mechanisms serve as different approximations of morality's ultimate function: finding mutually beneficial solutions to multi-agent decision problems (Levine, Chater, et al., 2024; André et al., 2022). By explicitly modeling both a rule's purpose and the generative mechanism that translates complex purposes into simple rules, we provide insight into how people reason about when rules apply and when they can be broken or updated.

Rules and Resource-Rational Moral Psychology Our results align with the "resource-rational contractualism" framework, which proposes that while direct negotiation would be ideal for moral decisions, people employ various approximations based on available cognitive resources (Lieder & Griffiths, 2020). The model demonstrates how rule interpretation can be understood as a resource-rational process, where people balance the computational costs of considering all possible exceptions against the benefits of more nuanced rule application. This connects to evidence that people use more resource-intensive, accurate processes rather than heuristic ones when moral stakes are high (Wu et al., 2024; Trujillo et al., 2024).

Community Context and Rule Understanding A key finding from our simulations is how identical actions can carry different moral weight depending on community composition. This aligns with cognitive frameworks that model communicative inferences through intuitive social theories (Frank & Goodman, 2012; Goodman & Frank, 2016; Degen, 2023). While previous work focused on dyadic communication, our results extend these approaches to community-directed rules, showing how people may reason about collective action and values.

Future Directions Important questions remain about the cognitive processes underlying rule interpretation. Given that people show variation in moral judgments (Curry, 2016; Barrett & Saxe, 2021; Graham et al., 2011), future work should investigate how our model captures different patterns of human moral reasoning. Recent computational work has begun characterizing these differences (Levine, Rottman, et al., 2020; Levine, Kleiman-Weiner, et al., 2024; Trujillo et al., 2024), suggesting potential extensions of our framework to account for individual variation in rule interpretation strategies.

Additionally, while we focused on a specific case study of grass-walking rules, the principles we uncovered—about how people reason about rules' functions, community needs, and appropriate exceptions—may extend to other domains of moral reasoning. Future work could examine how these mechanisms generalize across different types of moral rules and social contexts, particularly those involving more complex trade-offs between individual and collective utilities.

Acknowledgments

This project was supported by grants from the Templeton World Charity Foundation to S.L. LW and JBT acknowledge support from AFOSR Grant #FA95501910269, the MITIBM Watson AI Lab, MIT Quest for Intelligence, ONR Science of AI, and the Simons Center for the Social Brain.

References

André, J.-B., Debove, S., Fitouchi, L., & Baumard, N. (2022, May). *Moral cognition as a nash product maximizer: An evolutionary contractualist account of morality.* PsyArXiv. Retrieved from psyarxiv.com/2hxgu doi: 10.31234/osf.io/2hxgu

Barrett, H. C., & Saxe, R. R. (2021). Are some cultures more mind-minded in their moral judgements than others? *Philosophical Transactions of the Royal Society B*, *376*(1838), 20200288.

Curry, O. S. (2016). Morality as cooperation: A problem-centred approach. In *The evolution of morality* (pp. 27–51). Springer.

Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., ... others (2024). Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*(1), 519–540.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Fuller, L. L. (1958). Positivism and fidelity to law–a reply to professor hart. *Harv. L. Rev.*, *71*, 630.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Hart, H. L. A. (1958). Positivism and the separation of law and morals. In *Law and morality* (pp. 63–99). Routledge.

Klingefjord, O., Lowe, R., & Edelman, J. (2024). What are human values, and how do we align ai to them? *arXiv preprint arXiv:2404.10636*.

Kwon, J., Tan, Z.-X., Tenenbaum, J., & Levine, S. (2023). When it's not out of line to get out of line: The rules of rule-breaking. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (2024). *Resource-rational contractualism: A triple theory of moral cognition.* Brain and Behavioral Sciences.

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J. B. (2024). When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment. *Cognition*, *250*, 105790. doi: https://doi.org/10.1016/j.cognition.2024.105790

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. In *Proceedings of the national academy of sciences.*

Levine, S., Rottman, J., Davis, T., O'Neill, E., Stich, S., & Machery, E. (2020). Religious affiliation and conceptions of the moral domain. *Social Cognition*.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Misyak, J. B., & Chater, N. (2014). Virtual bargaining: a theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130487.

Portner, P. (2007). Imperatives and modals. *Natural language semantics*, *15*, 351–383.

Schauer, F. (1987). Precedent. *Stanford Law Review*, 571–605.

Trujillo, D., Zhang, M., Zhi-Xuan, T., Tenenbaum, J. B., & Levine, S. (2024). Resource-rational virtual bargaining for moral judgment: Towards a probabilistic cognitive model. *TopiCS in Cognitive Science*.

Wu, S. A., Ren, X., Gerstenberg, T., Choi, Y., & Levine, S. (2024). Resource-rational moral judgment. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Zhi-Xuan, T., Carroll, M., Franklin, M., & Ashton, H. (2024). Beyond preferences in ai alignment. *arXiv preprint arXiv:2408.16984*.