When Is It Acceptable to Break the Rules? Knowledge Representation of Moral Judgements Based on Empirical Data (Extended Abstract)

JAAMAS Track

Edmond Awad
University of Exeter
Exeter, UK
e.awad@exeter.ac.uk

Sydney Levine
Allen Institute for AI
Seattle, USA
smlevine@mit.edu

Andrea Loreggia
University of Brescia
Brescia, Italy
andrea.loreggia@gmail.com

Nicholas Mattei
Tulane University
New Orleans, USA
nsmattei@tulane.edu

Iyad Rahwan
Center for Humans & Machines, Max Planck Institute for Human Development
Germany
sekrahwan@mpib-berlin.mpg.de

Francesca Rossi
IBM Research
Yorktown Heights, NY, USA
francesca.rossi2@ibm.com

Kartik Talamadupula
Wand AI
Seattle, USA
kartik@wand.ai

Joshua Tenenbaum
Massachusetts Institute of Technology
Boston, USA
jbt@mit.edu

Max Kleiman-Weiner
University of Washington
Seattle, USA
maxkw@uw.edu

ABSTRACT

This paper explores how humans make contextual moral judgments to inform the development of AI systems capable of balancing rule-following with flexibility. We investigate the limitations of rigid constraints in AI, which can hinder morally acceptable actions in specific contexts, unlike humans who can override rules when appropriate. We propose a preference-based graphical model inspired by dual-process theories of moral judgment and conduct a study on human decisions about breaking the social norm of "no cutting in line." Our model outperforms standard machine learning methods in predicting human judgments and offers a generalizable framework for modeling moral decision-making across various contexts. This short paper summarizes the main findings of our paper published in the journal *Autonomous Agents and Multi-Agent Systems.* [2]

KEYWORDS

Moral decision-making, thinking fast and slow, human judgment

ACM Reference Format:
Edmond Awad, Sydney Levine, Andrea Loreggia, Nicholas Mattei, Iyad Rahwan, Francesca Rossi, Kartik Talamadupula, Joshua Tenenbaum, and Max Kleiman-Weiner. 2025. When Is It Acceptable to Break the Rules? Knowledge Representation of Moral Judgements Based on Empirical Data (Extended Abstract): JAAMAS Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

The increasing use of AI systems in real-world applications has raised significant concerns about their ethical alignment and behavior [10, 14, 16]. A key challenge is ensuring that AI systems operate within morally acceptable boundaries while avoiding issues like "specification gaming," where the system exploits loopholes in its design [1, 12, 15]. Traditional approaches to controlling AI behavior often rely on rigid, rule-based constraints. However, these can be either too restrictive—blocking reasonable actions—or too permissive, allowing harmful outcomes. In contrast, humans navigate such dilemmas with remarkable flexibility, intuitively overriding rules when context demands it. This human capacity for context-sensitive moral reasoning highlights the limitations of both static rule-based and reward specification approaches in AI [1, 5, 8].

To address this gap, we propose Scenario-Evaluation-Preference Networks (SEP-nets), an extension of CP-nets that allows for the representation of preferences not only over outcomes but also over decision contexts. Inspired by dual-process theories of moral cognition, SEP-nets model the two modes of human reasoning: fast, rule-based judgments (System 1) and slower, deliberative reasoning (System 2) [6, 7, 9, 17]. To validate our framework, we both collect data on and examine human moral judgments in the context of a simple yet illustrative social rule: "no cutting in line." This scenario reveals how people flexibly apply and override the rule based on contextual factors. By modeling these judgments using SEP-nets, we demonstrate improved prediction of human decisions compared to standard machine learning methods. Our work provides a promising foundation for building AI systems that can balance strict rule-following with the flexibility needed to navigate complex, real-world ethical dilemmas.

Model	Accuracy	F1	Precision	Recall	Time (ms)
RandomForest	0.7651 (0.0069)	0.7119 (0.0190)	0.7402 (0.0121)	0.6859 (0.0276)	303 (24)
XGBoost	**0.7870 (0.0227)**	0.7307 (0.0417)	0.7822 (0.0325)	0.6868 (0.0556)	109 (4)
Vorace	0.7166 (0.0091)	0.6620 (0.0178)	0.6692 (0.0212)	0.6550 (0.0152)	181955 (10173)
SVM	0.7115 (0.0192)	0.6493 (0.0203)	0.6704 (0.0103)	0.6298 (0.0299)	6157 (454)
SEP-Table	**0.7870 (0.0261)**	**0.7329 (0.0367)**	0.7817 (0.0354)	**0.6906 (0.0438)**	**22 (1)**
SEP-SVM	0.7834 (0.0248)	0.7224 (0.0340)	**0.7926 (0.0458)**	0.6654 (0.0424)	259 (17)

Table 1: Average performance on the test sets and the average training time of the different models in a 5-fold cross-validation, standard deviation in parentheses. Best performance in bold.

2 SEP-NETS: SCENARIOS, EVALUATION, AND PREFERENCE NETWORKS

SEP-nets (Scenarios, Evaluation, and Preference Networks) are a generalization of CP-nets designed to model context-aware decision-making and human-like moral reasoning [13]. Unlike CP-nets [4], which only handle preference variables, SEP-nets introduce three distinct types of variables: *Scenario Variables (SVs)* that define the context of the decision, *Evaluation Variables (EVs)* that model cognitive processes of evaluation, and *Preference Variables (PVs)* that capture the final preferences. This allows SEP-nets to represent a richer and more flexible decision-making process. The variable dependencies follow a three-level acyclic structure, where SVs have no parents, EVs depend on SVs or other EVs, and PVs can depend on any other variable. This layered structure models both fast, intuitive decisions (System 1) and slower, deliberative reasoning (System 2), inspired by dual-process theories of moral cognition [3, 11].

The semantics of SEP-nets define a preference order over *SEP-outcomes*, which are full assignments to all variables (SVs, EVs, and PVs). Two SEP-outcomes are comparable only if they share the same assignments for SVs and EVs. Preferences are then induced by a CP-net structure over the PVs. The process of identifying optimal outcomes follows three steps: (1) selecting a scenario by fixing the SVs, (2) setting the EVs according to evaluation functions that provide real-valued judgments, and (3) determining the most preferred assignment to the PVs according to the induced CP-net. This approach allows SEP-nets to model context-sensitive and ethically nuanced decision-making. When SVs and EVs are absent, SEP-nets reduce to classic CP-nets, demonstrating that SEP-nets are a strict extension of CP-nets. This richer formalism makes SEP-nets well-suited for modeling moral reasoning and context-dependent decision-making in AI systems.

3 DATA COLLECTION AND RESULTS

To test our model, we ran an experiment on Amazon MTurk. Informed consent was given by all participants and this study was approved by the Massachusetts Institute of Technology Institutional Review Board. Subjects were randomly assigned to one of three story contexts, in which subjects were asked to imagine that they were standing in line as a *deli* (12 scenarios), for a single-occupancy *bathroom* (7 scenarios), or at an *airport* security screening (6 scenarios). These contexts present the opportunity for someone to want to cut in line for a diverse range of reasons. We developed cases that manipulated (1) the amount of time by which the person cutting would delay the line, (2) the benefit that the person cutting would accrue by cutting, (3) the benefit that the people waiting would accrue by this person cutting, (4) the likelihood this particular scenario would happen at all.

To demonstrate its feasibility, two SEP-net variants were developed — SEP-SVM and SEP-Table — and tested on a binary prediction task. The goal was to model social behavior in a specific scenario, predicting whether an individual would allow another to cut in line based on contextual information (location and reason) and evaluations of welfare-related variables. The two SEP-nets differ in how evaluation is modeled: SEP-SVM employs Support Vector Machines (SVMs) to predict evaluation values, while SEP-Table uses empirical distributions derived from training data, partitioning evaluation values into quartiles. Both SEP-nets adopt a simple approach to construct the conditional preference table for the preference variable, using the most frequent preference in the training set.

To assess the performance of SEP-nets, they were compared with several machine learning models commonly used for binary classification, including XGBoost, Random Forest, VORACE (ensemble of neural networks), and a single SVM. All models were trained using 5-fold cross-validation, and their performance was evaluated using accuracy, precision, recall, F1-score, and training time. Table 1 reports the results that revealed that SEP-Table achieved the best overall performance, matching the state-of-the-art XGBoost in accuracy, while significantly outperforming it in training time. The SEP-SVM also achieved competitive performance, surpassing the single SVM and demonstrating the advantage of specialization within the SEP-net framework. The low variance observed in SEP-net results suggests their robustness and reliability in modeling context-dependent social decisions. This highlights the potential of SEP-nets to model complex human moral reasoning, capturing both intuitive (System 1) and deliberative (System 2) cognitive processes.

4 DISCUSSION

Our data collection and analysis highlight two central results. First, they provide evidence for our hypothesis about moral psychology, namely, that System 2 (outcome-based and agreement-based) reasoning is at play for our participants when deciding when to override rules. Second, they demonstrate that our novel formalism (SEP-nets) describes this process in a way that could be useful for enabling AI systems to understand the bounds of constraints. Additional results, details, and discussions can be found in the full version of our paper [2].

ACKNOWLEDGMENTS

Nicholas Mattei was supported by NSF Awards IIS-RI-2007955, IIS-III-2107505, IIS-RI-2134857, IIS-RI-2339880 and CNS-SCC-2427237.

REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *CoRR* abs/1606.06565 (2016). arXiv:1606.06565 http://arxiv.org/abs/1606.06565

[2] Edmond Awad, Sydney Levine, Andrea Loreggia, Nicholas Mattei, Iyad Rahwan, Francesca Rossi, Kartik Talamadupula, Joshua B. Tenenbaum, and Max Kleiman-Weiner. 2024. When is it acceptable to break the rules? Knowledge representation of moral judgements based on empirical data. *Auton. Agents Multi Agent Syst.* 38, 2 (2024), 35. https://doi.org/10.1007/S10458-024-09667-4

[3] Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jonathan Lenchner, Nick Linck, Andreas Loreggia, Keerthiram Murgesan, Nicholas Mattei, Francesca Rossi, et al. 2021. Thinking Fast and Slow in AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 15042–15046.

[4] C. Boutilier, R. Brafman, C. Domshlak, H.H. Hoos, and D. Poole. 2004. CP-nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements. *Journal of Artificial Intelligence Research* 21 (2004), 135–191.

[5] Jack Clark and Dario Amodei. 2016. Faulty reward functions in the wild. https://blog.openai.com/faulty-reward-functions. Accessed: Aug 1, 2023.

[6] Fiery Cushman. 2013. Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review* 17, 3 (2013), 273–292.

[7] Joshua David Greene. 2014. *Moral tribes: Emotion, reason, and the gap between us and them.* Penguin, London.

[8] HLA Hart. 1958. Positivism and the Separation of Law and Morals. *Harvard Law Review* 71 (1958), 593–607.

[9] Daniel Kahneman. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York.

[10] Michael Kearns and Aaron Roth. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design.* Oxford University Press, New York.

[11] Sydney Levine, Nick Chater, Joshua Tenenbaum, and Fiery Cushman. 2023. Resource-rational contractualism: A triple theory of moral cognition. (2023).

[12] Andrea Loreggia, Nicholas Mattei, Taher Rahgooy, Francesca Rossi, Biplav Srivastava, and Kristen Brent Venable. 2022. Making Human-Like Moral Decisions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 447–454. https://doi.org/10.1145/3514094.3534174

[13] Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K. Brent Venable. 2018. Preferences and Ethical Principles in Decision Making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 222. https://doi.org/10.1145/3278721.3278723

[14] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown, New York.

[15] Francesca Rossi and Nicholas Mattei. 2019. Building Ethically Bounded AI. In *Proc. of the 33rd AAAI(Blue Sky Track)*.

[16] Stuart Russell, Sabine Hauert, Russ Altman, and Manuela Veloso. 2015. Ethics of artificial intelligence. *Nature* 521, 7553 (2015), 415–416.

[17] Steven A Sloman. 1996. The empirical case for two systems of reasoning. *Psychological bulletin* 119, 1 (1996), 3.