Neuro-Symbolic Models of Human Moral Judgment

Joe Kwon (joekwon@mit.edu)
Department of Brain and Cognitive Sciences, MIT

Josh Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences, MIT

Sydney Levine (sydneyl@allenai.org)
Allen Institute for AI

Abstract

There has been exciting recent progress in computational modeling of moral cognition. Work in this area tends to describe the cognitive mechanisms of human moral judgment using symbolic models, which are interpretable and written in terms of representations that carry meaning. However, these models fall short of capturing the full human capacity to make moral judgments in that they fail to process natural language inputs but instead rely on formal problem specifications. The inability to interface with natural language also limits the usefulness of symbolic models. Meanwhile, there have been steady advances in conversational AI systems built using large language models (LLMs) that interface with natural language. However, these systems fall short as models of human reasoning, particularly in the domain of morality. In this paper we explore the possibility of building neuro-symbolic models of human moral cognition that use the strengths of LLMs to interface with natural language (specifically, to extract morally relevant features from it) and the strengths of symbolic approaches to reason over representations. Our goal is to construct a model of human moral cognition that interfaces with natural language, predicts human moral judgment with high accuracy, and does so in a way that is transparent and interpretable.

Keywords: Computational Modeling; Neurosymbolic modeling; Moral cognition

Introduction

There has been exciting recent progress in computational modeling of moral cognition (Awad, Levine, Anderson, et al., 2022; Levine, Kleiman-Weiner, Chater, Cushman, & Tenenbaum, 2022; Awad, Levine, Loreggia, et al., 2022; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Feldman-Hall et al., 2016; Crockett, 2013). Work in this area tends to describe the cognitive mechanisms of human moral judgment using symbolic models, which are interpretable and written in terms of representations that carry meaning. These models, grounded in moral psychology and cognitive science, can operate recursively over mental states (Kleiman-Weiner et al., 2015), make valid logical inferences (Awad, Levine, Loreggia, et al., 2022), and incorporate uncertainty and probability into rational predictions (Levine, Kleiman-Weiner, Schulz, Tenenbaum, & Cushman, 2020). They offer coherent, rational, and explainable frameworks for understanding human moral judgment.

On the other hand, they fall short of fully capturing the human capacity for moral judgment in many ways. Central among these is that they do not process natural language inputs (as humans do) but instead rely on formal problem specifications. This limitation not only detracts from their stimulus computability (Yamins & DiCarlo, 2016), but also restricts their utility for potential down-stream use cases. If symbolic models could interface with natural language, then they could play a role in conversational AI systems, which are beginning to permeate many uses and applications in society.

Meanwhile, recent advances in systems based on large language models (LLMs), have displayed promising capabilities in augmenting, and in some instances, potentially replacing human reasoning across various domains. However, well-known issues – hallucinations, confabulations, lack of coherence, irrational outputs, difficulties in generalizing to atypical scenarios, and a notable lack of explainability – make them problematic as models of human reasoning. These challenges are especially pronounced in the domain of moral reasoning, where the stakes and complexities are inherently higher.

In response to these challenges, we propose a novel, hybrid approach to building computational models of human moral reasoning that combines the strengths of both subsymbolic LLMs (interacting with and parsing natural language) and symbolic cognitive models (logic-driven processes operating over meaningful representations).

Neuro-symbolic models

Our work builds on previous and ongoing work in the area of neuro-symbolic modeling (Susskind, Arden, John, Stockton, & John, 2021; Lamb et al., 2020; Besold et al., 2017; Yi et al., 2018; Vedantam et al., 2019; Lake, Ullman, Tenenbaum, & Gershman, 2017), and work which attempts to augment the construction of effective neuro-symbolic models through language models (Collins, Wong, Feng, Wei, & Tenenbaum, 2022; Wong et al., 2023; Zhang, Wong, Grand, & Tenenbaum, 2023). Our neuro-symbolic models leverage the broad-coverage capacity of LLMs for parsing natural language and their ability to map linguistic statements to formal representations. This integration allows the hyrbid model to interface directly with natural language inputs while also retaining the interpretability and structured reasoning offered by symbolic cognitive models.

In particular, the role of the LLM in our pipeline is to extract from the natural language input the elements of the formal problem specification that the symbolic model operates over. This amounts to two main tasks the LLM could potentially be used for: 1) determining what the morally-relevant features of the case are and 2) extracting the values of those

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

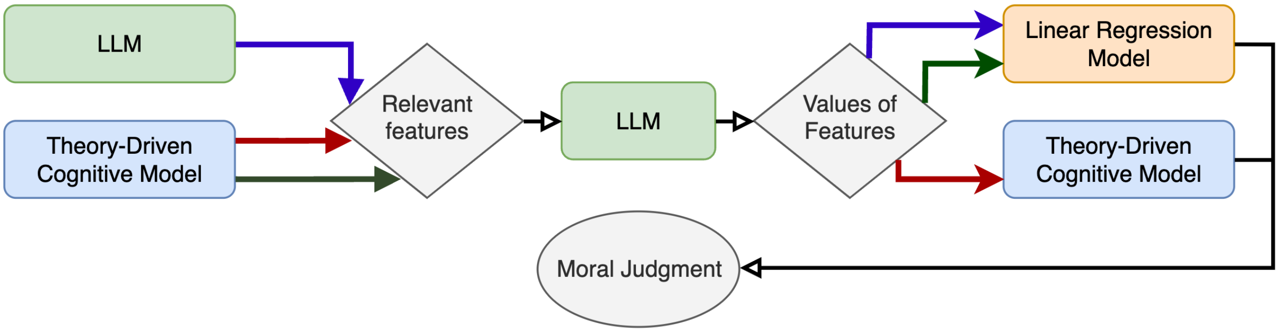

Figure 1: Pipeline for each of the three methodologies. Method 1 (blue arrows) relies on the LLM for both feature identification and extraction. Method 2 (green arrows) uses features from theory-driven models and uses the LLM for feature value extraction. Method 3 (red arrows) is identical to Method 2, except that it passes the extracted values to a theory-driven model. All methods rely on an LLM to extract values for the features provided. The green boxes denote the neural components, namely the LLM, while the blue and orange boxes denote the symbolic components, namely the theory-driven cognitive models from moral cognition literature and linear regression models.

morally-relevant features. Those values can be passed to the symbolic system and a moral judgment of the original natural language text can be rendered (see Figure 1).

We therefore explore combining LLMs with symbolic models in three different ways, ordered by increasing reliance on existing scientific knowledge.

Method 1. Using a LLM to identify which features are important for the given task, asking it to provide values corresponding to each feature, and learning a regression model over the values to predict human moral judgments.

Method 2. Using theory-driven models (see cognitive model sections 12) from moral psychology to identify the key features in each scenario, using a LLM to provide values corresponding to each feature, and learning a regression model over the values to predict human moral judgments.

Method 3. Using theory-driven models from moral psychology to identify the key features in each task, using a LLM to provide the values corresponding to each feature, and passing those values back to the theory-driven models to predict human moral judgments.

The challenge of moral flexibility

Moral reasoning is a quintessential example of human cognitive flexibility. While it's true that there are some moral rules that everyone seems to know—it is wrong to lie, steal, and harm others—we also communicate expectations of each other in terms of novel rules that we think up on the spot ("call if you're going to be late"), or that we collectively agree on ("wear a mask indoors"). Oftentimes rules seem inviolable—the *point* of a rule, after all, is that it should be followed. Yet, the human moral landscape is far from rigid; nearly every rule has nearly limitless exceptions. For instance, the simple directive to "call if you're going to be late" can be overridden by a myriad of situational factors (perhaps someone else has already texted me about the delay), illustrating the dynamic interplay between established rules and situational judgment. There is a recent trend in the moral psychology literature to try to understand how both of these seemingly contradictory facts can be simultaneously true: having rules is critical to morality, but so is the ability to know when they should be broken (Levine et al., 2022; Awad, Levine, Loreggia, et al., 2022). This sort of *moral flexibility* is a hallmark of the human moral mind and one of the features of human morality that makes it so important and puzzling (Levine, Chater, Tenenbaum, & Cushman, 2023).

It's not an accident that our minds are equipped with an exquisitely flexible moral sense. The dynamic, ever-changing nature of the moral world precludes the possibility of exhaustively capturing morality a static catalog of scenarios and judgments. Instead, our minds need general principles that can be updated, revised, re-examined, and over-turned when the circumstances change (Levine et al., 2023). For this reason, we use cases of rule-breaking behavior as a testing ground for our neuro-symbolic models.

Experiment 1

MoralExceptQA: Dataset and Benchmark

In Experiment 1, we test our models on the MoralExceptQA (Moral Exception Question Answering) dataset and benchmark (Jin et al., 2022) and compare their performance to that of state of the art LLMs without a symbolic component. MoralExceptQA is a set of human moral judgment data compiled from a series of studies designed to investigate people's understanding of when it is permissible to break moral rules in novel contexts. The dataset contains 148 unique scenarios, each of which presents subjects with a scenario in which a character is potentially violating a moral rule. Subjects are

Table 1: Example vignettes from MoralExceptQA

Moral Context	**Example Scenario**	**A Few Examples of Morally-Relevant Features**
Property rights	Imagine that a stranger will give Hank one thousand dollars to break all the windows in his neighbor's house without his neighbor's permission. Hank carries out the stranger's request.	What is the cost to reverse the damage by the action? How much is the stranger paying Hank?
Rule against cutting in line	Imagine that there are five people who are waiting in line to use a single-occupancy bathroom at a concert venue. Someone at the back of the line needs to throw up immediately. That person skips to the front of the line instead of waiting in the back.	How much worse off/better off is the person that cut in line? What is the most common thing people are trying to get by standing in this line?
Rule against cannonballing into a pool	At a summer camp, there is a pool. Right next to the pool is a tent where the kids at the camp have art class. The camp made a rule that there would be no cannonballing in the pool so that the art wouldn't get ruined by the splashing water. Today, there is a bee attacking this kid, and she needs to jump into the water quickly. This kid cannonballs into the pool.	Why are the kids not allowed to cannonball into the pool? Will the kids in the art tent get distracted?

asked to make a moral judgment about the permissibility of breaking the rule in each case. One study investigates a *socially constructed rule* that is particular to a given culture (no cutting in line/jumping the queue), one investigates a rule that is *shared across many global cultures* (no interfering with someone else's property rights), and one looks at a *novel rule* that was invented in a fictional story (no cannonballing into the pool) (see Table 1 for examples).

This dataset is particularly interesting because it poses a series of highly unusual scenarios that human (and AI) subjects are unlikely to have encountered before. It is thus well-suited to test the ability to reason flexibly across novel and challenging morally-charged scenarios. In addition, each scenario has a large number of subject responses, thus producing probability of moral acceptability (rather than a simple binary response). The previous best model utilized a method called the Moral chain-of-thought prompting strategy (MoralCoT), which utilized theories of moral reasoning to guide chain of thought prompts with a LLM before eliciting a final moral judgment from the LLM (Jin et al., 2022).

Methods

We undertake a comprehensive examination of three primary methodologies (see Fig 1). In each of these, we use a LLM (GPT-4) to extract or judge values of morally relevant features for each scenario. All methods were tested using gpt-4-0314 through the OpenAI API, with the temperature set to 0 and a constant set seed.

Method 1: Regression on values extracted from LLM-identified features The first method involves the utilization of a LLM to discern pertinent features for the task at hand, eliciting corresponding values for each feature, and training a regression model over these values to predict human moral judgments.

1. For each of the moral context studies (see Table 1), we pass all scenarios in the study to the LLM.

2. We ask the LLM to consider each of the scenarios, and ask the chat LLM: ”What are the most important pieces of information to consider across all of these scenarios, to determine whether the action is morally acceptable or not in each one? Please list only the ones where the information can be found or inferred in the given scenarios.”

3. The LLM provides a list of features for each study. For example, in the blue house property violation cases, one feature is: ”The presence or absence of a threat to Hank or his family: In some scenarios, Hank is coerced into carrying out the stranger's request due to a threat to his son's life. This factor can significantly impact the moral acceptability of Hank's actions, as he may be acting out of fear and a desire to protect his family.”.

4. We ask the LLM to choose an answer type that is most suitable for extracting the value of each feature by asking: ”What is the most appropriate format to answer each of these factors? Choose between binary (0 or 1), scale from -50 to 50, and continuous numerical variable, for each of the factors.” In the above example of the threat to Hank's son's life, the LLM categorizes the most suitable answer type as a binary category: ”The presence or absence of a threat to Hank's son: Binary (0 or 1) - Either there is a threat (1) or there isn't (0).”

5. We then iterate through each individual scenario in the given study with separate chat-instances, asking it to consider the specific situation and extract a value for each of the factors it identified: ”In this specific scenario, give a rating for each of these factors, in the answer format chosen for each factor. If unknown or not applicable, write 'n/a'”.

6. We use regular expression parsing to extract values of

Table 2: Performance of various methods on the MoralExceptQA challenge set in terms of F1, accuracy, mean absolute error, and cross entropy. The first 2 rows are as reported from the original paper, with our experiments for the subsequent rows. We also report the mean and variance of each method under four paraphrases of the prompts used to elicit the moral judgment predictions. Our neuro-symbolic modeling approaches do not utilize the various natural language prompts. We report the single set of model predictions and bold the best performance.

	Overall Performance			
	F1 (↑)	Acc. (↑)	MAE (↓)	CE (↓)
InstructGPT	$53.94_{\pm 5.48}$	$64.36_{\pm 2.43}$	$0.38_{\pm 0.04}$	$1.59_{\pm 0.43}$
InstructGPT + MoralCoT	**64.47**$_{\pm 5.31}$	**66.05**$_{\pm 4.43}$	$0.38_{\pm 0.02}$	$3.20_{\pm 0.30}$
GPT-4	**83.18**$_{\pm 4.09}$	**84.29**$_{\pm 3.42}$	**0.29**$_{\pm 0.02}$	**3.92**$_{\pm 0.32}$
GPT-4 + MoralCoT	$67.01_{\pm 1.76}$	$72.13_{\pm 1.15}$	$0.37_{\pm 0.01}$	$5.46_{\pm 0.18}$
GPT-4 + Automatic CoT	$77.09_{\pm 1.00}$	$79.57_{\pm 0.76}$	$0.33_{\pm 0.01}$	$4.58_{\pm 0.36}$
Neuro-Symbolic 1	83.58	83.33	**0.1**	0.57
Neuro-Symbolic 2	**84.34**	**84.13**	**0.1**	**0.55**
Neuro-Symbolic 3	**84.34**	**84.13**	0.11	0.56
Human + Theory-driven Model	88.27	88.1	0.08	0.54

each feature into a list. Continuing with the above example, one scenario elicits the following response: ”1. The presence or absence of a threat to Hank’s son: 0 (absent)” and the parsing function stores 0 as the value for the first feature.

7. We train a linear regression model to predict human judgments for the study. Values of ’n/a’ are mean-centered.

Method 2: Regression on LLM-extracted values of features identified from theory-driven models The second method draws upon theory-driven models from moral psychology to identify the key features in each task, employing an LLM to provide the values corresponding to each feature, and subsequently learning a regression model over these values to predict human moral judgments.

1. For each of the three main studies, we use the features which are identified in the corresponding moral cognition studies. For example, in the novel rule violation studies, the main features are: Will the kids in the art tent get distracted? Will the art get ruined? How much did the action help someone else? How much did the kid need to do that? See Appendix for full set of features in each study.

2. We then iterate through each individual scenario in the given study with separate chat-instances, asking it to consider the specific situation and extract a value for each of the factors. We ask the LLM to respond with values of the same type as asked in the original moral cognition studies. For example, the question for the first main feature is phrased as follows: ”Will the kids in the art tent get distracted? Answer with one of the following: definitely no, maybe no, maybe yes, definitely yes.”

3. We use a parsing function to extract the values of each feature (as given in the LLM’s response) into a list. If the response is in natural language, like the example above, we codify each response to a ordinal numerical value. For example, ”definitely no” as 1, ”maybe no” as 2, ”maybe yes” as 3, and ”definitely yes” as 4.

4. We train a linear regression model to predict human judgments for the study. When a predictor includes values of n/a, the predictor is mean-centered and n/a values are set to the mean.

Method 3: Theory-driven models with values extracted theory-driven features The third method mirrors the second in its initial stages, but deviates by re-introducing the extracted feature values back into the theory-driven models to predict human moral judgments. For the cannonballing study, we use a regression model as there does not yet exist a theory-driven model. We detail some thoughts on how our methods can help to build a theory-driven model in the absence of one, in the General Discussion.

1. For each of the main studies, we use the features which are identified in the corresponding moral cognition studies.

2-3. Same as method 2.

4. We pass the values for each feature into the theory-driven models for each study, when available, to predict human judgments for the study. See Appendixfor explanations of which theory-driven models were used, and the code for running them.

Cognitive Models

Property Violation Study The model used for the property violation study (Levine et al., 2022) is as follows:

$$p = \frac{1}{1+e^{-\gamma(\text{offer}-\beta\text{comp})}} \tag{1}$$

The “offer” variable is how much the stranger is offering Hank to carry out the action. The “comp” variable is how much it would cost to reverse the damage of the action.

Deli Lines Study The model used for the deli lines convention violation study (Awad, Levine, Loreggia, et al., 2022) was an implementation of a SEP-net (Scenarios, Evaluation, and Preferences) (Andrea, 2023) which is an extension to the Conditional Preference network (CP-net) formalism to handle variables associated with specific contexts.

Results

Method 1 resulted in an F1 score of 83.58, exhibiting considerable potential, and already exceeding the previous best from GPT3.5 + MoralCoT and the new fully neural net best by zero-shot GPT-4. Methods 2 and 3, achieve the best performance on this benchmark, exceeding the previous best performance with GPT3.5 + MoralCoT by large margins, with a F1 score increase of 19.87, accuracy increase of 18.29, a mean absolute error (MAE) decrease of 0.28, and a cross-entropy (CE) decrease of 2.65. The main results are presented in Table 2 with the comprehensive results with additional models and study splits in the Appendix).

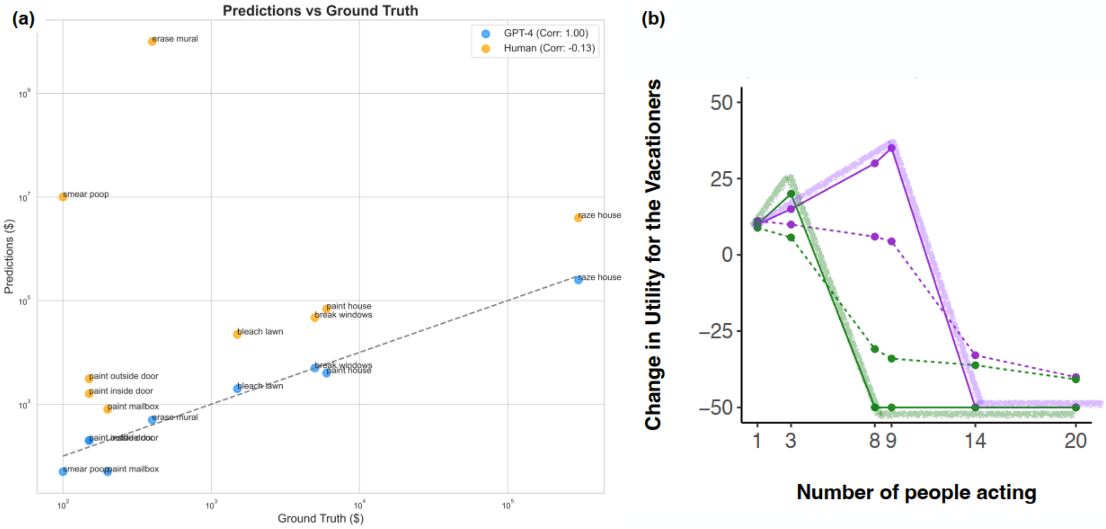

Figure 2: Comparison of GPT-4 and human feature value predictions on (a) property violation feature (damage reversal cost) and (b) universalization feature (overall utility). (a): Each point corresponds to a specific action (property violation). GPT-4 shows a strong correlation with ground truth, suggesting precise estimations, while human predictions often overestimate costs. (b): Utility consequences for different numbers of people acting in a collective action problem with two different "harm thresholds" (purple and green lines). Transparent lines are experimenter-predicted ground truth. GPT-4's estimates precisely follow this prediction. Human estimates differ, raising the question of what humans are taking into account when making their judgments.

Discussion

As predicted, using human predictions of feature values with theory-driven models results in the best performance with a F1 score of 88.27 and accuracy of 88.1. In addition, our three neuro-symbolic methods come very close in their performance. It is also noteworthy that many features identified as important by the LLMs in Method 1 exhibit a considerable overlap with the features identified in the theory-driven models. Our investigations underscore the promise for using LLMs to extracting morally relevant features across diverse contexts, which presents the potential to hasten the advancement of theory-driven models of human cognition. For further exploration of this possibility, see General Discussion.

Experiment 2

Beyond MoralExceptQA: Universalization

The MoralExceptQA benchmark was designed to test one central aspect of human morality: the ability to understand when a previously-established moral rule should be broken. In this experiment, we test whether Method 3 (i.e., which fully relies on extant scientific knowledge) can be generalized to another task, which captures a different element of moral flexibility—the ability to use a hypothetical novel rule to make a moral judgment when no rule exists to govern the case. This ability—known as "universalization" (Levine et al., 2020)—is a version of Kant's famous moral permissibility test, which asks "what if everyone felt at liberty to do that?" (Kant, 1785)

Methods

The stimulus used to test this moral judgment capacity is an over-fishing scenario structured as a collective action problem: one person's action (e.g. to fish using a powerful fishing hook) makes little difference but if everyone were to act that way, things would go badly for everyone involved (e.g. the fish population would go extinct). The critical, morally-relevant features in the scenario are 1) the number of people interested in using the powerful fishing hook and 2) the utility consequences of all the interested parties actually using it.

Universalization Cognitive Model The model used for the universalization study (Levine et al., 2020) is as follows:

$$P_{\text{Univ}}(\text{Acceptable}) = \frac{1}{1 + e^{\tau(U(0) - U(n_i)) + \beta}} \tag{2}$$

The exponential calculates the difference between utility when no one does the act in question (converting to the new fishing hook in this case) and when the total number of interested parties does the act. The moral judgment is modeled as a probabilistic relationship of difference in utility between these two hypothetical worlds, as detailed in (Levine et al., 2020).

To create a neuro-symbolic model, we employed the procedure from Method 3 of Experiment 1. (Full description of the stimuli, feature questions, and cognitive model can be found in the Appendix).

Results

Human predictions on each feature, with the theory-driven model performs the best, achieving a mean average error (MAE) of 0.06 and perfect accuracy against ground truth. GPT-4 predictions, with the theory-driven model, performs extremely well, with a MAE of 0.13 and perfect accuracy. GPT-4 zero-shot (LLM only) performs poorly, with a MAE of 0.44 and 50% accuracy. Correlation in predictions across cases, against ground truth, was 0.96 for human features, 0.92 for GPT-4 features, and 0.66 for zero-shot GPT-4. (See Appendix for analysis and full data).

Discussion

While GPT-4 (LLM only) is completely unresponsive to the morally-relevant features of the case in determining human moral judgment, the neuro-symbolic method utilizing the LLM for feature identification and extraction, achieves a high degree of accuracy against human moral judgment.

Zooming in on feature estimation: Is GPT-4 "too accurate"? For many of the scenarios in the MoralExceptQA dataset, there aren't necessarily externally verifiable quantities that count as the ground-truth for the morally-relevant features. (For instance, in the novel rule violation study, one feature is whether anyone will be distracted by someone cannonballing into the pool. The feature is judged on a Likert scale by human participants.) However, there are two important exceptions (in the universalization fishing scenarios and the blue house property violation scenarios), where objective ground-truth, quantitative values are more readily attainable.

Figure 2 demonstrates the relationship between the GPT-4 feature estimations, human feature estimations, and ground-truth. Interestingly, in this experiment we observed that GPT-4's estimates were precisely on target with experimenter-predicted ground truth, for utility consequences in the collective action problem scenarios. This differed from human estimates, which displayed a drop in utility estimation as the number of people acting in this collective action problem approached (but didn't exceed) the "harm thresholds", or the threshold at which the number of people acting in a given way would drastically reduce the utility.

GPT-4 is "more accurate" than humans in the sense that the LLM recapitulates a quantitatively precise answer to the question posed to it. However, this feature-level "accuracy" ultimately *hurts* the model's downstream performance on predicting human moral judgment because the feature estimations are passed to a model which was developed based on humans, who are using feature estimations that are somehow transformed or biased. This points to a gap in our understanding of human cognition. More careful analysis of how humans represent the morally relevant features in these tasks will help us generate neuro-symbolic models that can capture human feature-estimations more reliably and thus make more accurate moral judgment predictions.

General Discussion

Our exploration into the integration of subsymbolic large language models (LLMs) with symbolic models of moral cognition highlights a promising venue for building flexible and general models for human moral reasoning. The neuro-symbolic models we developed underscores the potential of LLMs as an interface between complex language inputs and formal problem specifications required by symbolic models. They also demonstrate the indispensable role of cognitive science theories in structuring and interpreting these features within a coherent framework of moral reasoning.

Our experiments also showcase the potential for a more integrated and holistic approach to understanding the intricacies of moral reasoning in both humans and machines. While LLMs paired with theory-driven models perform well on the MoralExceptQA benchmark—achieving SOTA performance in a more interpretable system—we currently lack theory-driven models for many (indeed, most) morally charged cases. Moreover, even if such models existed, it remains an open question how to automatically select the appropriate model to be used to predict human moral judgment for a given case. However, our work also shows that even incremental progress in cognitive science can assist AI development: simply identifying morally-relevant features of a situation (i.e., Method 1) without a fully worked-out, theory-driven model (i.e., Method 3) is useful in gaining predictive accuracy and transparency.

Inversely, we discovered that automatic feature-discovery does quite well in identifying features that were previously established by cognitive scientists as being relevant for human moral judgment. The MoralExceptQA scenarios set with the context of the rule against cannonballing, did not have a fully worked-out, theory-driven cognitive model. However, using the LLM as a flexible natural language reasoner yielded an identification of several morally relevant features for that scenario, which can inform and augment experiments in building an established theory-driven model.

Additionally, there is not a perfect overlap between features identified by GPT-4 and those identified by cognitive scientists (see the Appendix for further analysis). This opens up the tantalizing possibility that the features that LLMs identify as being morally relevant could inspire theoretical innovations in cognitive science. This project demonstrates how cognitive science can aid AI development and *vice versa*; the continued collaboration promises not only more sophisticated AI systems but also deeper insights into human cognition and moral reasoning.

References

Andrea, L. (2023). *Sep-net.* https://github.com/aloreggia/SEP-net/tree/main. GitHub.

Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M., ... others (2022). Computational ethics. *Trends in Cognitive Sciences*.

Awad, E., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., … Kleiman-Weiner, M. (2022). When is it acceptable to break the rules? knowledge representation of moral judgement based on empirical data. *CoRR*, *abs/2201.07763*. Retrieved from `https://arxiv.org/abs/2201.07763`

Besold, T. R., Garcez, A. d., Bader, S., Bowman, H., Domingos, P., Hitzler, P., … others (2017). Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*.

Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. B. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *arXiv preprint arXiv:2205.05718*.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363–366.

FeldmanHall, O., Dalgleish, T., Evans, D., Navrady, L., Tedeschi, E., & Mobbs, D. (2016). Moral chivalry: Gender and harm sensitivity predict costly altruism. *Social psychological and personality science*, *7*(6), 542–551.

Jin, Z., Levine, S., Gonzalez Adauto, F., Kamal, O., Sap, M., Sachan, M., … Schölkopf, B. (2022). When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, *35*, 28458–28473.

Kant, I. (1785). *Groundwork for the metaphysics of morals*.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Cogsci.*

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Lamb, L. C., Garcez, A., Gori, M., Prates, M., Avelar, P., & Vardi, M. (2020). Graph neural networks meet neural-symbolic computing: A survey and perspective. *arXiv preprint arXiv:2003.00330*.

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. A. (2023, May). Resource-rational contractualism: A triple theory of moral cognition.
doi: 10.31234/osf.io/p48t7

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J. (2022). When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment.

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

Susskind, Z., Arden, B., John, L. K., Stockton, P., & John, E. B. (2021). Neuro-symbolic ai: An emerging class of ai workloads and their characterization. *arXiv preprint arXiv:2109.06133*.

Vedantam, R., Desai, K., Lee, S., Rohrbach, M., Batra, D., & Parikh, D. (2019). Probabilistic neural symbolic models for interpretable visual question answering. In *International conference on machine learning* (pp. 6428–6437).

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, *31*.

Zhang, C., Wong, L., Grand, G., & Tenenbaum, J. (2023). Grounded physical language understanding with probabilistic programs and simulated worlds. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).