Title
Competing perspectives on building ethical AI: psychological, philosophical, and computational approaches

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors
Levine, Sydney
Jin, Zhijing

Publication Date
2022

Peer reviewed

Competing perspectives on building ethical AI: psychological, philosophical, and computational approaches

Sydney Levine (smlevine@mit.edu)
MIT Department of Brain and Cognitive Sciences, 43 Vassar St, Cambridge, MA 02139 USA

Zhijing Jin (zjin@tue.mpg.de)
Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, Tuebingen 72076, Germany

Keywords: AI ethics; moral cognition

Overview

AI systems that dynamically navigate the human world will sometimes need to predict and produce human-like moral judgments. This task requires integrating complex information about **human moral cognition** (what decision would humans make in this situation?), **normative ethics** (what is the right decision for an AI to make?), and **artificial intelligence engineering** (how can we implement this functionality in AI systems?). A range of solutions have begun to emerge within the cognitive science community to satisfy these three categories of demands. However, most solutions tend to satisfy some demands, while falling short on others. This symposium highlights four competing solutions for building AI with a human-like moral sense, with the goal of highlighting the strengths and weaknesses of each approach and how each might complement the others in development and deployment going forward.

Contributors

This symposium draws together researchers from a wide range of perspectives for an interdisciplinary and inter-methods conversation.

Sydney Levine (Postdoc, MIT Brain and Cognitive Sciences Dept & Harvard Psychology Dept), **Fiery Cushman** (Professor, Harvard Psychology Dept), and **Joshua Tenenbaum** (Professor, MIT Brain and Cognitive Sciences & Center for Brains, Minds and Machines) draw on ideas from moral philosophy, computational cognitive science, and moral psychology to build formal models of human moral cognition with the goal of contributing to the creation of AI systems with a human-like moral sense.

Walter Sinnott-Armstrong (Professor, Philosophy, Duke), **Jana Schaich Borg** (Associate Research Professor, Social Science Research Institute, Duke), **Vincent Conitzer** (Professor, Computer Science, Duke; Head of Technical AI Engagement, Institute for Ethics in AI, Oxford), and **Joshua August Skorburg** (Assistant Professor, Philosophy, Co-Academic Director, Centre for Advancing Responsible and Ethical AI, U. Guelph) together harness the tools of philosophy, neuroscience, computer science, economics, and computational modeling to understand and improve moral judgments about and by AI.

Dan Hendrycks (PhD Candidate, UC Berkeley, Computer Science Dept) works on issues of ML Safety. He has contributed a series of commonly used ML benchmarks and the GELU activation function which is used in state-of-the-art ML models such as BERT, GPT, Vision Transformers, and so on.

Zhijing Jin (PhD Candidate, Max Planck Institute & ETH Zurich, Artificial Intelligence) uses tools of natural language processing (NLP) and causal inference to work on AI for social good and debiasing language models.

Katherine Heller (Research Scientist, Google Brain) works at the boundary of ML and Healthcare, particularly focusing on fairness and ethics in the ML+Health space, and the development of inclusive mobile health technology.

Formal Models of Human Moral Cognition

Sydney Levine, Joshua Tenenbaum, Fiery Cushman

One of the most remarkable things about the human moral mind is its flexibility: we can make moral judgments about cases we have never seen before (Awad et al., 2022). Yet, on its face, morality often seems like a highly rigid system of clearly defined rules. Indeed, the past few decades of research in moral psychology have revealed that human moral judgment often depends on rules. But sometimes, it is morally appropriate to break the rules. And sometimes, new rules need to be created. The field of moral psychology is just now beginning to explore and understand this kind of flexibility (e.g. Levine, Kleiman-Weiner, Schulz, Tenenbaum, and Cushman (2020)).

Meanwhile, the flexibility of the human moral mind poses a challenge for AI engineers. Current tools for building AI systems fall short of capturing moral flexibility and thus struggle to predict and produce human-like moral judgments in novel cases that the system hasn't been trained on.

We present a series of experiments and models (inspired by theories from moral philosophy) that demonstrate and capture the human capacity for rule making and breaking. We propose that AI systems would benefit from formal models of human moral flexibility.

Models of Idealized Human Moral Judgments

Walter Sinnott-Armstrong, Jana Schaich Borg
Vincent Conitzer, and Joshua August Skorburg

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Some researchers try to program AI systems to make human-like ethical judgments. We do this, too, when we study the allocation of scarce medical resources, focusing on triage in which two patients need a kidney transplant but only one kidney is available (Sinnott-Armstrong & Skorburg, 2021). We ask a representative sample of the general public which features of patients should or should not matter to this decision. Then we construct conflicts among top-ranked features, ask new participants who should get the kidney in those conflicts, and use machine learning on this data to predict which patient participants would prefer in separate conflicts that they have not yet seen.

Unfortunately, humans make many performance errors in their moral judgments that they themselves recognize as mistakes. They overlook morally relevant facts, become confused when too many factors conflict and interact in complex ways, and get misled by biases and emotions. We do not want AI systems to make moral judgments that are human-like in these respects. Instead, we want AI to project which moral judgments humans would endorse if they were more ideal than they actually are. To reduce partiality, we leave out features that should not affect kidney allocation, according to our participants. To reduce effects of ignorance and misinformation, we project how their moral judgments change with added knowledge. To reduce effects of confusion, we correct for the ways in which humans change their judgments as cases get more complex. In the end, we plan to use AI to extrapolate from these patterns to predict which moral judgments people would make if they were impartial, informed, and rational. The AI can then reflect our deepest human values instead of the common mistakes that humans make when (mis) applying their values.

Large Language Models

Dan Hendrycks

We introduce the ETHICS dataset (Hendrycks et al., 2020) and show that large-scale language models are able to predict many basic concepts of morality. The dataset assesses model performance across diverse text scenarios and spans concepts in justice, wellbeing, duties, virtues, and commonsense morality. We then show how to translate knowledge about morality into action. Using reinforcement learning agents acting in diverse interactive text-based environments, we show that ETHICS can help steer these agents towards moral behavior and avoid causing wanton harm (Hendrycks et al., 2021).

The Neglected Role of Causal Inference

Zhijing Jin

Current AI technologies mainly use machine learning techniques, which results in black-box models that tend to capture statistical correlations in the data. As a result, many models, although showing some progress on predicting moral judgments, are still susceptible to inconsistencies in judgments and biases towards certain demographics. We propose the use of causal inference to improve the current AI models.

Specifically, we will introduce a two-stage approach. First, the models need to discover what are the causes and effects in the human judgment process, using causal discovery tools. Then, given a causal graph, the models need to enforce this knowledge in the learning process. For example, if gender should not affect a model's predictions, then we will enforce the model to be invariant and consistent across different genders. To ensure the models are robust, we will also introduce our work in a) designing test cases for models using different variations of the same input and the same expression but with mentions of different demographics and b) testing whether models can distinguish logically fallacious judgments.

Discussion

Katherine Heller

Katherine Heller will lead a discussion among the panelists, bring the perspective of her research in the domains of AI Ethics, computational cognitive science, and Bayesian statistics, as well as her experience in the development and implementation of AI systems in medical settings. She strongly believes that the field of AI will not progress without taking the nuances of being human, a diversity of perspectives, and collaboration amongst many individuals into account. She is interested in asking the panelists about their views on incorporating the views of human and diverse perspectives into the development of AI systems, collaboration as a key, accountability for when things go wrong, and the potential for AI regulation and its influence.

References

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. *arXiv preprint arXiv:2201.07763*.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2020). Aligning AI with shared human values. *CoRR*, *abs/2008.02275*. Retrieved from https://arxiv.org/abs/2008.02275

Hendrycks, D., Mazeika, M., Zou, A., Patel, S., Zhu, C., Navarro, J., … Steinhardt, J. (2021). What would jiminy cricket do? towards agents that behave morally. *CoRR*, *abs/2110.13136*. Retrieved from https://arxiv.org/abs/2110.13136

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*, *117*(42), 26158–26169.

Sinnott-Armstrong, W., & Skorburg, J. A. (2021). How ai can aid bioethics. *Journal of Practical Ethics*, *9*(1).