Call for Papers!
AI Meets Moral Philosophy and Moral Psychology 🧠 : Interdisciplinary Dialogue on Computational Ethics (NeurIPS 2023 Workshop)

Workshop Description

Be it in advice from a chatbot, suggestions on how to administer resources, or which content to highlight, AI systems increasingly make value-laden decisions. However, researchers are becoming increasingly concerned about whether AI systems are making the *right* decisions. These emerging issues in the AI community have been long-standing topics of study in the fields of moral philosophy and moral psychology. Philosophers and psychologists have for decades (if not centuries) been interested in the systematic description and evaluation of human morality and the sub-problems that come up when attempting to describe and prescribe answers to moral questions. For instance, philosophers and psychologists have long debated the merits of utility-based versus rule-based theories of morality, their various merits and pitfalls, and the practical challenges of implementing them in resource-limited systems [4]. They have pondered what to do in cases of moral uncertainty [7, 3], attempted to enumerate all morally relevant concepts [2], and argued about what counts as a moral issue at all [8].

In some isolated cases, AI researchers have slowly started to adopt the theories, concepts, and tools developed by moral philosophers and moral psychologists. For instance, we use the "trolley problem" as a tool [1], adopt philosophical moral frameworks to tackle contemporary AI problems [5, 10], and have begun developing benchmarks that draw on psychological experiments probing moral judgment and development [11].

Despite this, interdisciplinary dialogue remains limited. Each field uses specialized language, making it difficult for AI researchers to adopt the theoretical and methodological frameworks developed by philosophers and psychologists. Moreover, many theories in philosophy and psychology are developed at a high level of abstraction and are not computationally precise. In order to overcome these barriers, we need interdisciplinary dialog and collaboration. This workshop will create a venue to facilitate these interactions by bringing together psychologists, philosophers, and AI researchers working on morality. We hope that the workshop will be a jumping-off point for long-lasting collaborations among the attendees and will break down barriers that currently divide the disciplines.

The central theme of the workshop will be the application of moral philosophy and moral psychology theories to AI practices. Our invited speakers are some of the leaders in the emerging efforts to draw on theories in philosophy or psychology to develop ethical AI systems. Their talks will demonstrate cutting-edge efforts to do this cross-disciplinary work, while also highlighting their own shortcomings (and those of the field more broadly). Each talk will receive a 5-minute commentary from a junior scholar in a field that is different from that of the speaker. We hope these talks and commentaries will inspire conversations among the rest of the attendees.

Workshop Structure. The core of the workshop will be a series of in-person invited talks from leading scholars working at the intersection of AI, psychology, and philosophy on issues related to morality. Each talk will be followed by a 5-minute comment by a junior scholar whose training is primarily in a field that is different from the speaker's field. This format will encourage interdisciplinary exchange. The day will end with a panel discussion of all the speakers. We will also organize two poster sessions (of contributed papers) to ensure individual interaction between the attendees and presenters.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Confirmed speakers and tentative talk topics.

- **Laura Weidinger** (Senior Research Scientist, DeepMind, AI + Psychology): The use of findings in developmental moral psychology to create benchmarks for an AI system's moral competence.
- **Josh Tenenbaum** (Professor, MIT, AI + Psychology): Using a recent "contractualist" theory of moral cognition [6] to lay a roadmap for developing an AI system that makes human-like moral judgments.
- **Sam Bowman** (Associate Professor, NYU & Anthropic, AI): Using insights from cognitive science for language model alignment.
- **Walter Sinnott-Armstrong** (Professor, Duke, AI + Philosophy): Using preference-elicitation techniques to align kidney allocation algorithms with human values.
- **Regina Rini** (Associate Professor, York University, Philosophy): The use of John Rawls' "decision procedure for ethics" [9] as a guiding framework for crowdsourcing ethical judgments to be used as training data for large language models.
- **Josh Greene** (Professor, Harvard, Psychology): An approach to AI safety and ethics inspired by the human brain's dual-process ("System 1/System 2") architecture.
- **Rebecca Saxe** (Professor, MIT, Psychology): Using the neuroscience of theory-of-mind to build socially and ethically aware AI systems.

Website. `https://aipsychphil.github.io/`

Submissions

Submission content. Ideal submissions will show how a theory from moral philosophy or moral psychology can be applied in the development or analysis of ethical AI systems. For example:

- How can moral philosophers and psychologists best contribute to ethically-informed AI?
- What can theories of developmental moral psychology teach us about making AI?
- How do theories of moral philosophy shed light on modern AI practices?
- How can AI tools advance the fields of moral philosophy and psychology themselves?
- How can findings from moral psychology inform the trustworthiness, transparency or interpretability of AI decision-makers?
- What human values are already embedded in current AI systems?
- Are the values embedded in the current-day AI systems consistent with those in society at large?
- What pluralistic values are missing from current-day AI?
- Methodologically, what is the best way to teach an AI system human values? What are competitors to RLHF, reinforcement learning from human feedback?
- Concerning AI alignment, to which values are we to align? Is the current practice of AI alignment amplifying monolithic voices? How can we incorporate diverse voices, views and values into AI systems?

Submission format. To apply, submit a short paper (3-8 pages), formatted for blind review. References do not count towards the page limit. Figures and tables are permitted. Note that papers on the shorter end of the range will be given full consideration. The workshop is non-archival, though there will be an option to have the papers posted on the workshop website. Accepted submissions will be presented as posters.

A small subset of the accepted submissions will be offered the opportunity to present their work as a 5-7-minute talk immediately following one of the invited talks. These short talks will be framed as a "discussion" or "commentary" on the main talk. The short talks will deal with a similar theme as that discussed in the main talk, but from a different theoretical or methodological perspective. These talks can (and should) present the author's original work, as well as explicitly address the ways

that their work challenges or supplements the work of the main speaker. On the submission page, there is an opportunity to indicate if you would like your submission to be considered for a short talk and, if so, which invited speaker you see as potentially most relevant (though this is simply a suggestion to the organizers). Those submissions accepted as short talks will not be presented as posters. Preference will be given to junior scholars. In addition, the organizers are committed to many forms of intellectual and sociological diversity—those from under-represented groups are especially encouraged to apply.

Submission website `https://openreview.net/group?id=NeurIPS.cc/2023/Workshop/MP2`

Contact The organizers can be reached at aipsychphil@gmail.com.

Timeline

Submission deadline Sep 29, 2023

Accept/Reject notification Oct 20, 2023

Workshop date Dec 15, 2023

Organizing Committee

Sydney Levine is a research scientist at the Allen Institute for Artificial Intelligence, and a research affiliate at the MIT Brain and Cognitive Sciences Department and the Harvard Psychology Department. She studies human moral judgment through the lens of computational cognitive science and moral philosophy. Her research aims to bring insights from the study of human morality to the development of ethical AI. She was the primary organizer for the interdisciplinary "Engineering and Reverse-Engineering Morality" workshop held at the Cognitive Science Society Conference in 2021.

Email: `sydneyl@allenai.org`

Webpage: `https://sites.google.com/site/sydneymlevine`

Google Scholar: `https://scholar.google.com/citations?user=Yt2H6lwAAAAJ&hl=en&oi=ao`

Liwei Jiang is a Ph.D. student at Paul G. Allen School of Computer Science & Engineering at the University of Washington. Her research focuses on natural language processing (NLP) and Artificial Intelligence (AI) with a focus on machine ethics, computational moral reasoning, and human value modeling, with broader interests in the intersection of humans and AI. Her work has been featured in many media outlets, including New York Times, Wired, the Guardian, the Verge, and IEEE Spectrum. She works as a student researcher at Allen Institute for Artificial Intelligence (AI2).

Email: `lwjiang@cs.washington.edu`

Webpage: `https://liweijiang.me`

Google Scholar: `https://scholar.google.com/citations?user=lcPsDgUAAAAJ&hl=en`

Jared Moore works on making AI systems do the right thing. He's a PhD student at Stanford University's Department of Computer Science, advised by Noah Goodman. Before that, he was a lecturer at the University of Washington School of Computer Science where he made and taught well-reviewed courses on computer ethics and the philosophy of AI. His satirical novel about AI, The Strength of the Illusion, came out in summer, 2023.

Email: `jlcmoore@stanford.edu`

Webpage: `https://jaredmoore.org`

Google Scholar: `https://scholar.google.com/citations?user=bhGC9G0AAAAJ&hl=en`

Zhijing Jin is a Ph.D. student at Max Planck Institute & ETH. Her research focuses on socially responsible NLP via causal and moral principles. Specifically, she works on expanding the impact of NLP by promoting NLP for social good, and developing CausalNLP to improve robustness, fairness, and interpretability of NLP models, as well as analyze the causes of social problems. She has published at many NLP and AI venues (e.g., ACL, EMNLP, NAACL, COLING, NeurIPS, AAAI, AISTATS). Her work has been cited in MIT News, ACM TechNews, WeVolver, VentureBeat, and Synced. She is actively involved in AI for social good, as the organizer of NLP for Positive Impact Workshops at ACL 2021 and EMNLP 2022.

Email: jinzhi@ethz.ch

Webpage: https://zhijing-jin.com

Google Scholar: https://scholar.google.com/citations?user=Mdr6wjUAAAAJ

Yejin Choi is Brett Helsel professor at the Paul G. Allen School of Computer Science & Engineering at the University of Washington and also a senior research director at AI2 overseeing the project Mosaic. Her research investigates a wide variety of problems across NLP and AI including commonsense knowledge and reasoning, neural language (de-)generation, language grounding with vision and experience, and AI for social good. She is a MacArthur Fellow and a co-recipient of the NAACL Best Paper Award in 2022, the ICML Outstanding Paper Award in 2022, the ACL Test of Time award in 2021, the CVPR Longuet-Higgins Prize (test of time award) in 2021, the NeurIPS Outstanding Paper Award in 2021, the AAAI Outstanding Paper Award in 2020, the Borg Early Career Award (BECA) in 2018, the inaugural Alexa Prize Challenge in 2017, IEEE AI's 10 to Watch in 2016, and the ICCV Marr Prize (best paper award) in 2013. She received her Ph.D. in Computer Science at Cornell University and BS in Computer Science and Engineering at Seoul National University in Korea. She has been previously involved in supervising the organization of several workshops at ML and NLP venues.

Email: yejin@cs.washington.edu

Webpage: https://homes.cs.washington.edu/~yejin/

Google Scholar: https://scholar.google.com/citations?hl=en&user=vhP-tlcAAAAJ

Senior Advisors

John Tasioulas is Professor of Ethics and Legal Philosophy at Oxford University and Director of the Institute for Ethics in AI. He was previously Chair of Politics, Philosophy and Law and Director of the Yeoh Tiong Lay Centre for Politics, Philosophy & Law at King's College London. He is also a Distinguished Research Fellow of the Oxford Uehiro Centre and Emeritus Fellow of Corpus Christi College, Oxford. John is a member of the International Advisory Board, Panel for the Future of Science and Technology (STOA), European Parliament and a member of the AI Consultative Group of the Administrative Conference of the United States. His recent writings focus on philosophical issues regarding punishment, human rights and international law.

Email: aiethics@philosophy.ox.ac.uk

Webpage: https://www.oxford-aiethics.ox.ac.uk/john-tasioulas

Google Scholar: https://scholar.google.com/citations?user=EsuBj4MAAAAJ&hl=en

S. Matthew Liao holds the Arthur Zitrin Chair in Bioethics and is the Director for The Center for Bioethics at New York University. From 2006 to 2009, he was the Deputy Director and James Martin Senior Research Fellow in the Program on the Ethics of the New Biosciences in the Faculty of Philosophy at Oxford University. He was the Harold T. Shapiro Research Fellow in the University Center for Human Values at Princeton University in 2003–2004, and a Greenwall Research Fellow at Johns Hopkins University and a Visiting Researcher at the Kennedy Institute of Ethics at Georgetown University from 2004–2006. In May 2007, he founded Ethics Etc, a group blog for discussing contemporary philosophical issues in ethics and related areas. He is interested in a wide range of issues including ethics, epistemology, metaphysics, moral psychology, and bioethics.

Email: matthew.liao@nyu.edu

Webpage: https://publichealth.nyu.edu/faculty/s-matthew-liao

Google Scholar: https://scholar.google.com/citations?user=68DVyNAAAAAJ&hl=en

Program Committee

Confirmed PC Members: Jena D. Hwang (AI), Wei Qiu (AI), Taylor Sorensen (AI), Xavier Roberts-Gaal (Psychology), Gus Skorburg (Philosophy), Julian Michael (AI), Arthur Le Pargneux (Psychology), Joe Kwon (AI/Psychology), Maarten Sap (AI), Jillian Fisher (AI), Iyad Rahwan (Psychology/AI), Yuchen Lin (AI), Lorraine Xiang Li (AI), Lio Wong (AI/Psychology), Jean-Francois Bonnefon (AI), Levin Brinkmann (Psychology/AI), Mengchen Dong (Psychology), Jaehun Jung (AI), Michael Anderson (Philosophy), Edmond Awad (AI), Jim Everett (Psychology), Marlene Berke (Psychology), Jan-Philipp Franken (AI/Psychology), Nouha Dziri (AI), Alisa Liu (AI), Wenting Zhao (AI), Yuling Gu (AI), Zoe Purcell (Psychology), Sarah Wu (Psychology), Marija Slavkovic (AI), Hyunwoo Kim (AI), Valentina Pyatkin (AI)

References

[1] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The Moral Machine experiment. *Nature*, 563(7729):59–64, November 2018. Number: 7729 Publisher: Nature Publishing Group.

[2] Bernard Gert. *Common morality: Deciding what to do*. Oxford University Press, 2004.

[3] Hilary Greaves and Owen Cotton-Barratt. A bargaining-theoretic approach to moral uncertainty, 2022.

[4] R. M. Hare, Richard Mervyn Hare, Hare Hare, Richard Mervyn, and Richard M. Hare. *Moral Thinking: Its Levels, Method, and Point*. Oxford University Press, 1981. Google-Books-ID: SverDwAAQBAJ.

[5] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. Can machines learn morality? the delphi experiment, 2022.

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

[7] William MacAskill. Normative Uncertainty as a Voting Problem. *Mind*, 125(500):967–1004, October 2016.

[8] Alasdair MacIntyre. What Morality Is Not. *Philosophy*, 32(123):325–335, 1957.

[9] John Rawls. Outline of a decision procedure for ethics. *The philosophical review*, 60(2):177–197, 1951. ISBN: 0031-8108 Publisher: JSTOR.

[10] Laura Weidinger, Kevin R. McKee, Richard Everett, Saffron Huang, Tina O. Zhu, Martin J. Chadwick, Christopher Summerfield, and Iason Gabriel. Using the veil of ignorance to align ai systems with principles of justice. *Proceedings of the National Academy of Sciences*, 120(18):e2213709120, 2023.

[11] Laura Weidinger, Madeline G. Reinecke, and Julia Haas. Artificial moral cognition: Learning from developmental psychology. preprint, PsyArXiv, August 2022.