Benchmarked Ethics: A Roadmap to AI Alignment, Moral Knowledge, and Control

Aidan Kierans
aidan.kierans@uconn.edu
University of Connecticut
Storrs, Connecticut, USA

ABSTRACT

Today's artificial intelligence (AI) systems rely heavily on Artificial Neural Networks (ANNs), yet their black box nature induces risk of catastrophic failure and harm. In order to promote verifiably safe AI, my research will determine constraints on incentives from a game-theoretic perspective, tie those constraints to moral knowledge as represented by a knowledge graph, and reveal how neural models meet those constraints with novel interpretability methods. Specifically, I will develop techniques for describing models' decision-making processes by predicting and isolating their goals, especially in relation to values derived from knowledge graphs. My research will allow critical AI systems to be audited in service of effective regulation.

CCS CONCEPTS

• **Computing methodologies** → *Control methods*; **Knowledge representation and reasoning**; *Philosophical/theoretical foundations of artificial intelligence.*

KEYWORDS

alignment, knowledge graphs, interpretability, auditing, control

ACM Reference Format:
Aidan Kierans. 2023. Benchmarked Ethics: A Roadmap to AI Alignment, Moral Knowledge, and Control. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3600211.3604764

1 COMPLETED WORK

Growing concerns about the AI alignment problem have emerged in recent years, with previous work focusing mostly on (1) qualitative descriptions of the alignment problem; (2) attempting to align AI actions with human interests by focusing on value specification and learning; and/or (3) focusing on either a single agent or on humanity as a singular unit. However, the field as a whole lacks a systematic understanding of how to specify, describe and analyze misalignment among entities, which may include individual humans, AI agents, and complex compositional entities such as corporations, nation-states, and so forth. Prior work on controversy in computational social science offers a mathematical model of contention among populations (of humans). In our paper "Quantifying Misalignment Between Agents" [2], my collaborators and I adapt this contention model to the alignment problem, and show how viewing misalignment can vary depending on the population of agents (human or otherwise) being observed as well as the domain or "problem area" in question. Our model departs from value specification approaches and focuses instead on the morass of complex, interlocking, sometimes contradictory goals that agents may have in practice.

2 FUTURE WORK

In order to achieve my goal of promoting verifiably safe AI, I will (1) Extend my existing work on measuring alignment to verify it in simulations; (2) Construct a knowledge graph (KG) to represent claims and arguments from moral philosophy; and (3) Connect patterns in ANN weights and structures to embedded reward expectations. These projects will each produce tools for analyzing and/or improving ANNs during their creation. Together, they will allow researchers to teach AI systems human-like moral intuitions via (2), relate those intuitions to actions in a training environment by interpreting the ANNs via (3), and compare the exhibited AI values to those of humans in order to quantify its alignment via (1). Since my approach is multi-disciplinary, the projects are ordered according to the rate of progress in the relevant disciplines, such that the first will not be outdated before it can combine with the last.

2.1 Extend alignment work

I will first extend my existing work [2], which reframes misalignment as a pairwise function applicable to an arbitrary number of parties on a per-issue basis. This framing provides a much-needed structure for analyzing realistic multi-agent scenarios, as opposed to scenarios common in existing research, which frequently assume incorrectly that humans have homogenous interests and/or that there is only one AI agent. I am currently extending this work by applying it in multi-agent simulation environments to verify the functional applicability of this framework. I will test the framework's applicability and value by evaluating the misalignment scores it produces under several complex multi-agent scenarios, in environments that incentivize some or all of them to variously cooperate and compete.

2.2 Build Moral Knowledge Graph

I will utilize a partially-automated, human-in-the-loop approach to construct the world's first comprehensive knowledge graph (KG) representing human beliefs about morality. To the best of my knowledge, such a KG will be the first resource for machine-readable

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604764

moral philosophy data. While the Allen Institute for AI's research prototype Delphi is trained to mimic human ethical norms, it was trained on datasets sourced from crowdwork and mostly unfiltered internet data and does not include a curated philosophy database [1]. As input, I will scrape content from the Stanford Encyclopedia of Philosophy (SEP), a detailed, reliable, and high-quality philosophy reference work, for entries that mention "ethics," "morality," or related words. Initially, I will use entry titles as entities and infer relations between them by using natural language processing to extract key relational phrases. I will also represent sections of entries as entities with a hierarchical relationship to main entries. Once the KG coherently represents ideas from the SEP, I will utilize active learning and manual, expert verification to improve the coverage and accuracy of the representations. By utilizing human-in-the-loop annotations with the help of volunteers recruited at university philosophy departments and online philosophy communities, the resulting model would yield both higher accuracy and broader coverage. In addition to accelerating the timeline for reviewing the KG, this crowdsourcing effort would ensure the reviewers' diversity of backgrounds, perspectives, and areas of expertise, thus improving the resulting fairness of the KG. The resulting moral philosophy KG would contain ethical stances and supporting arguments relevant to human decision-making, resulting in an excellent basis for training ANNs to model and reference moral views and intuitions. I will test this by comparing the predicted values and uncertainty outputs of an AI trained on this KG to those of human samples in the moral psychology literature and to the AI2 moral reasoning engine Delphi.

2.3 Synthesize with Interpretability

Finally, I will link AI models' actions to their incentives. Building on existing interpretability methods, I will first verify my hypothesis that instrumental goals have latent representations in their weights and/or structures. For example, I would locate which neurons and pathways in an ANN trained to maximize a video game score have the strongest correlation with collecting in-game coins. I will extend this method to isolate representations of how AI models meet their incentives, and how we can tune them to favor some goals over others. Each stage of my research will yield useful tools for AI researchers, engineers and policymakers: (1) During the testing phase of AI development, researchers and engineers will be able to use my alignment framework to quantify, reduce, and mitigate misaligned goals before and during deployment. Likewise, the ability to quantify an AI agent's alignment based on its incentives will support regulators and policymakers in evaluating mission critical systems and mitigating the risks of inadvertently creating broadly misaligned AI. (2) Training or fine-tuning an ANN to respect human moral concerns will be significantly easier with access to a comprehensive KG of ethics literature. Whether engineers are creating language models or agents that interact with the real world, penalizing morally objectionable, questionable, repulsive and/or ambiguous outputs will be so practical that it could be required for large projects. (3) The tools I create will allow direct modification of an AI agent's priorities by embedding morality, which will streamline the ability to align AI with human moral intuitions. The foundation for measuring risks posed by ANNs established by parts (1) and (2) of my research will assist AI developers to quickly hone in on a system's biggest moral and practical risks. The fine-grained control that my interpretability research will yield will support engineers in making minimal, targeted interventions, allowing models to be aligned in real-world industry applications quickly and without sacrificing performance. By including diverse stakeholders in the creation of the world's first moral KG, this project will also hold AI fairness implications. Overall, the project will yield important insights into promoting multi-agent cooperation in RL models, improving AI truthfulness and fairness while reducing biases, and adherence to specific moral values in general AI agents. Modern ANNs are plagued by uncertainty in terms of both latent knowledge and value, and their alignment to human interests, goals and values. My research will enable quantifying AI alignment in realistic, meaningful terms; create an unprecedented resource for machine-accessible moral philosophy knowledge in the form of a KG, enabling ANN incorporation of human values; and unlock greater ANN understanding and alignment by connecting outputs to specific internal representations.

3 CONCLUSION

My research will yield straightforward yet invaluable benefits by connecting state-of-the-art AI alignment with in-depth contemporary philosophical understanding. My proposed realistic framework for measuring alignment along with an accessible resource for moral philosophy will enable straightforward measurement of ANNs' potential harms. The ability to make targeted changes to ANNs will reduce harm and create social value in a meaningful, achievable manner.

REFERENCES

[1] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574* (2021).

[2] Aidan Kierans, Hananel Hazan, and Shiri Dori-Hacohen. 2022. Quantifying Misalignment Between Agents. Presented at NeurIPS ML Safety Workshop 2022.