Heuristical Moral Decision-Making with Weak Supervision

***Keywords*—** moral machines, weak supervision, ethics

Motivation

As the widespread application of artificial intelligence (AI) systems grows, so does the discovery of serious fairness and bias issues in AI applications, from face recognition to the hiring process. We can expect algorithms to be increasingly faced with morally ambiguous decisions: for instance, should a driverless car should swerve into oncoming traffic to avoid a jaywalker?

Technology companies, governments, and all AI practitioners in society are faced with a common problem: how do we easily build and maintain algorithms that make moral decisions, given the wide variety of applications and moral considerations that exist?

Prior approaches have used survey data to develop ethical models that mimic popular moral preferences. But to avoid issues of bias and misinformation, I seek instead to develop a framework for quickly generating training data for moral dilemmas based on a set of adjustable moral heuristics, determined by ethical or legal principles.

Background

The problem of moral AI has been approached in two primary ways: first, as the classification problem of discriminating between ethical and unethical decisions; second, as a computational social choice problem in which the moral opinions of real human beings are modeled and aggregated into an ethical determination through a set of game-theoretic voting rules or utility calculations. Both of these approaches rely on empirical data, either to model the moral preferences of individual voters or to train a classifier. For example, the Moral Machine project collected human judgments of autonomous vehicle behavior in certain moral dilemmas for driverless cars, including the jaywalking example [1].

Though these are promising steps toward artificially intelligent moral decision-making, they rely on a democratized approach to ethics that comes with distinct limitations. In the absence of high quality crowd-sourced data indicating moral preferences, practitioners often rely on Amazon Mechanical Turk, as in [2], or another less desirable survey method. In many cases, surveyed voters are not likely to have appropriate domain expertise to produce high quality ground truth data.

Proposition

In many cases, it may be more beneficial to apply simple, *a priori* ethical or legal principles dictated by a regulator or concerned practitioner. I propose to develop a framework for learning moral preferences from a limited, overlapping set of ethical heuristics. Specifically, I will use the data programming paradigm suggested by [3] to design a weakly supervised generative model for ethical decisions.

Fundamentally, I hypothesize that there are latent moral principles in any given set of legal or ethical rules that can be discovered and used to produce labeled training data for a moral machine. These abstract moral principles can be identified by generalization over a set of moral heuristics provided by domain experts and used to improve the ethical decision-making of intelligent machines.

Validation

I will use data from the Moral Machine platform to validate my weakly supervised approach to ethical decision-making against previous, strongly-supervised approaches. Assume that Moral Machine users were relatively well-informed and that the choices made in each scenario are morally optimal on average. If this is the case, how well does a model trained by my heuristics-based, generative approach match hu-

man ethical choices relative to prior approaches? The same approach can be applied to other, similar publicly available datasets, including the kidney exchange problem [2].

I also expect to formally prove the theoretical properties of my framework, and the computational efficiency of any new algorithms I develop.

Methods

1. **Generative Model.** The first phase of research will be the development of the generative model, which integrates noisy ethical rules to produce training labels. Besides developing algorithms for the generalization of moral principles, this phase will involve defining a general framework for stating ethical rules programmatically, using major moral preferences (e.g. a preference for young lives over older ones) evident in the Moral Machine dataset as sample heuristics. An interface for adjusting the intensity of heuristics will also be designed.

2. **Discriminative Model.** In the second phase, a model will be developed to generalize the moral heuristics formulated in the first phase. Algorithmically, the discriminative model is much simpler, but the performance of the discriminative model depends greatly on the amount of unlabeled data available, but many ethical problems occur rarely. It will be necessary to develop a justifiable method for generating hypothetical moral dilemmas for a given ethical problem. There is also a key opportunity for model interpretability: this phase will also investigate ways to represent moral generalizations in a meaningful way so that they are useful to human users for both problem insight and external applications.

3. **Validation & Extension.** In the final phase, the finished pipeline will be validated on the Moral Machine dataset and investigate any abstract moral patterns found during the generalization step for new insight. The model will be applied to new test case (e.g. the kidney exchange problem) to assess its external validity. Some thought will be

Intellectual Merit

This proposal constitutes an advancement both in the field of ethical AI and in the field of weakly supervised learning. Ethical AI stands to gain from the addition of a more *a priori* approach to the field of empirically trained moral machines. In the process of designing generative algorithms for this specific problem, I expect to develop new methods for label generation that will be useful in similar cases.

Broader Impact

Fairness, ethics, accountability, and transparency are growing and vital issues in the field of artificial intelligence. AI has the potential to extraordinarily benefit society, but also has the potential to do great harm if used incautiously.

The primary benefit of a weakly supervised framework for ethical decision-making is to make moral AI better and easier to implement for a growing industry. The more AI is used for important decision-making, the more ethical concerns matter. The ability to intelligently combine multiple, possibly contradictory, moral heuristics, rather than expend money and effort on potentially unreliable survey data, will be invaluable for organizations implementing ethical systems at scale.

[1] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J. F. Bonnefon, and I. Rahwan. The Moral Machine experiment. *Nature*, 563(7729):59–64, 11 2018.

[2] R. Freedman, J. Schaich Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer. Adapting a Kidney Exchange Algorithm to Align with Human Values. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, pages 115–115, New York, New York, USA, 2018. ACM Press.

[3] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment*, volume 11, pages 269–282. Association for Computing Machinery, 11 2017.