An Empirical Approach to Capture Moral Uncertainty in AI

Andreia Martinho
Delft University of Technology
Delft, The Netherlands
a.m.martinho@tudelft.nl

Maarten Kroesen
Delft University of Technology
Delft, The Netherlands
m.kroesen@tudelft.nl

Caspar Chorus
Delft University of Technology
Delft, The Netherlands
c.g.chorus@tudelft.nl

ABSTRACT

As AI Systems become increasingly autonomous they are expected to engage in complex moral decision-making processes. For the purpose of guidance of such processes theoretical and empirical solutions have been sought.

The underlying idea which motivates the theoretical debates that have dominated the literature is that conceptual agreement regarding the moral machinery of artificial agents should precede design endeavors. This led to a renewed interest in normative ethics that spanned several domains of knowledge and research. Prospects were made for moral machines based on moral theories such as deontology and consequentialism [1, 9, 10] but the problem of moral disagreement between competing moral theories and conflicting moral judgments was never surmounted [3]. A solution that has been advanced in the literature is to design AI Systems to be fundamentally uncertain about morality [3, 5]. Decisions made by these systems within the realm of moral uncertainty would be based on the assumption that there is no certainty about which moral theory is correct and therefore weights should be ascribed to different moral theories [3]. A particular theoretical framework for decision-making under moral uncertainty developed by William MacAskill has recently been outlined within the domain of AI morality by Kyle Bogosian [3, 7].

The empirical attempts to address moral reasoning in AI Systems rely on the idea that moral rules found in human morality should be reflected in AI moral reasoning. These rules would be empirically elicited and embedded in the systems. Empirical approaches feature however underlying convergence strategies that undervalue moral heterogeneity. Several studies reported moral preferences revealed by participants in the context of moral dilemmas and emphasis is often placed on the dominant trends as potential guidelines for AI moral reasoning [2, 4].

In this research we integrate both theoretical and empirical lines of thought to address the matters of moral reasoning in AI Systems. We reconceptualize the metanormative framework for decision-making under moral uncertainty [3, 7] using discrete choice analysis techniques and we operationalize it through a latent class choice model. The key assumption of these models is that a number of classes featuring different preferences exist within a population albeit each class is internally relatively homogeneous [6].

The discrete choice analysis-based formulation of the metanormative framework is theory-rooted and practical as it captures moral uncertainty through a small set of latent classes.

To illustrate our approach we conceptualize a society in which AI Systems are in charge of making policy choices. In the proof of concept two AI systems make policy choices on behalf of a society but while one of the systems uses a baseline moral certain model the other uses a moral uncertain model. It was observed that there are cases in which the AI Systems disagree about the policy to be chosen which we believe is an indication about the relevance of moral uncertainty.

The main novelty in this research is the operationalization of the re-conceptualized metanormative framework through discrete choice analysis. We acknowledge that our re-conceptualization fails to take into account the richness and subtleties of the work developed originally by MacAskill [7, 8] yet opening an avenue for further research that accommodates its various extensions. Finally our proof-of-concept also requires further research on the meaning and practical implications of moral uncertainty in artificial decision-making.

KEYWORDS

Artificial Intelligence, Morality, Moral Uncertainty, Metanormative Theory

ACM Reference Format:
Andreia Martinho, Maarten Kroesen, and Caspar Chorus. 2020. An Empirical Approach to Capture Moral Uncertainty in AI. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Societyety (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3375627.3375805

REFERENCES

[1] Michael Anderson and Susan Leigh Anderson. 2011. *Machine ethics.* Cambridge University Press.
[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59.
[3] Kyle Bogosian. 2017. Implementation of Moral Uncertainty in Intelligent Machines. *Minds and Machines* 27, 4 (2017), 591–608.
[4] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.
[5] Miles Brundage. 2014. Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence* 26, 3 (2014), 355–372.
[6] William H Greene and David A Hensher. 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological* 37, 8 (2003), 681–698.
[7] William MacAskill. 2014. *Normative uncertainty.* Ph.D. Dissertation. University of Oxford.
[8] William MacAskill. 2016. Normative uncertainty as a voting problem. *Mind* 125, 500 (2016), 967–1004.
[9] Thomas M Powers. 2006. Prospects for a Kantian machine. *IEEE Intelligent Systems* 21, 4 (2006), 46–51.
[10] Wendell Wallach, Colin Allen, and Iva Smit. 2008. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society* 22, 4 (2008), 565–582.

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02.
https://doi.org/10.1145/3375627.3375805