Estimating Weights of Reasons Using Metaheuristics: A Hybrid Approach to Machine Ethics[1]

Benoît Alcaraz, Aleks Knoks, David Streit[2]

The central aim of the interdisciplinary field of machine ethics is to design artificial agents that are able to act in ethically acceptable ways. This aim can be achieved, it seems, only if there is a way to meet the challenge of specifying a way of acquiring and representing *normative information* that allows for machine implementation.

Within machine ethics, one can distinguish between three families of approaches to meeting this challenge [1], with their own advantages and pitfalls. First, there are top-down approaches that encode normative information in a symbolic formalism. The main advantage of these approaches is that symbolic representations have a clear intuitive meaning, and that systems that operate on them result in decisions that are transparent and intelligible. However, the designers of top-down systems are forced to encode the normative information by hand, foreseeing all the myriads of ways in which contextual factors might have to be taken into account if the system is to deliver the correct decision. Second, there are bottom-up approaches that use machine learning techniques. Their main advantage is that they allow for complex normative information to be obtained via training, without any need to encode it by hand. The main pitfall here is that it is not clear what normative information the system has actually learned, and how the decisions that it delivers can be made intelligible [2]. Third, there are hybrid systems that aim to combine the advantages of top-down and bottom-up approaches, while avoiding their pitfalls. But while the idea of combining normative information represented in symbolic form with machine learning is very appealing, the field of machine ethics is still far away from converging on a single approach that would hold promise for broad applications.

Against this background, we propose a new hybrid approach that draws its motivation from an influential philosophical (metaethical) account, or informal model, of the structure of morality. According to this account, an action's deontic status – whether it's permissible, required, or forbidden – in any given situation is determined on the basis of the "normative weights" of a designated class of considerations usually called "normative reasons". We first provide a formal characterization of this account, drawing on recent work in formal argumentation. We then use the resulting formal model to develop a system that uses a genetic algorithm to estimate the normative weights of reasons on the basis of a given set of cases for which the morally correct outcomes are known. The weight estimates can then be used to determine the deontic statuses of actions in new cases. To the best of our knowledge, this is the first application of the metaethical account to the concerns of machine ethics.

[1] Forthcoming in: Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES2024), ACM Press.

[2] University of Luxembourg, email: firstname.lastname@uni.lu

We evaluate the framework and test it, amongst other things, for its predictive accuracy, scalability, and resistance to noisy training data. We show that under reasonable circumstances our framework is highly ($\geq 95\%$) accurate. Even when the training data only makes up a small percentage of all possible cases, it remains reasonably accurate ($\geq 80\%$). Additionally, noisy training data does not have a large impact on the accuracy of our framework and our framework can even accommodate highly noisy, and even inconsistent data.

In our talk, we will present these results as well as some recent novel empirical results.

References

[1] Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7:149–155, 2005.

[2] Ilaria Canavotto and John Horty. Piecemeal knowledge acquisition for computational normative reasoning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 171–180, New York, NY, USA, 2022. Association for Computing Machinery.