From Algorithmic Black Boxes to Adaptive White Boxes: Declarative Decision-Theoretic Ethical Programs as Codes of Ethics

Martijn van Otterlo
Department of Cognitive Science and Artificial Intelligence
Tilburg University, The Netherlands
m.vanotterlo@uvt.nl
http://martijnvanotterlo.nl

Ethical challenges of artificial intelligence (AI) are rising as technological advances are widely spread [4, 11]. In addition to critiques from legal and sociology scholars who study the influence and regulation of algorithms, nowadays AI researchers themselves are actively involved by creating AI technology that is intrinsically responsible, transparent and especially explainable [3, 10]. In this paper I introduce a novel way to formalize ethical decision making based on principles of reinforcement learning [12] in a practical, computational logic to help build understandable AI systems that are value aligned [6].

Declarative decision-theoretic ethical programs (DDTEPs) [5] declaratively specify (and solve) ethical decisions of intelligent systems modeled as decision-theoretic problems, based on the probabilistic programming language DT-PROBLOG [7]. Solutions are computed by considering all possible worlds modeled by the program. The general idea is to formalize what is known explicitly in the model, and use reasoning to compute optimal decisions. DDTEPs fit into logical approaches for ethical (or: value-driven) reasoning [2] but also relational reinforcement learning [9] and provides novel opportunities for explanation-focused computations [8] by reasoning over the logical parts of the model.

A (partial) toy example of a DDTEP for a self-driving car consists of a decision to either run_into_wall (killing the passenger) or a collision (killing a pedestrian). We can specify percepts for what is in front of the car and a rule that says what happens when a collision is made. The utility function defines the value of each outcome, making the optimal value −30 (amounting to kill the passenger by steering away).

```
(action) run_into_wall; collision.
(percepts) in_front_of_car(a), baby(a).
(rules) kill(X) :- in_front_of_car(X), collision.
(values) utility(run_into_wall, -30).
utility(kill(X), -20) :- pedestrian(X).
utility(kill(X), -40) :- baby(X).
```

DDTEPs prove successful for toy ethical domains [1], but can generally be applied to any kind of ethical reasoning where (some) domain knowledge is available. Partially observable contexts can model information gathering actions where the AI can ask humans what current values and norms are, thereby reaching value alignment. Specifying human values for a DDTEP comes with choices,

* This paper was recently published at the Int. Conf. on AI, Ethics and Society [5]
such as that a baby is worth more than a pedestrian (or even that they appear on the same scale). Such choices can be dependent on social, cultural and other factors as the large-scale experiment the Moral Machine \(^1\) aims to investigate. Similarly, DDTEPs support learning such that parameters (rewards, utilities, probabilities) may be induced from existing data or by observing humans. This also means that the structure of the ethical reasoning style may be relatively stable (e.g. the rules) but values may vary from context to context (cf. [5, 11]).

DDTEPs open up the black box of algorithms and make decision logic transparent while still allowing for machine learning to fill in additional details from data. This general pattern is a solution to value alignment in AI systems in complex domains: i) formalize existing norms and values transparently into a DDTEP, and ii) finetune parts of the program on data. Formalisms like DDTEPs will provide the technical means to design responsible AI systems that can reason about their ethical decisions, but more importantly provide the mechanisms to compute explanations [8] for those decisions, and explain their behaviors to humans, thereby obtaining value alignment and trust [10].

References


\(^1\) http://moralmachine.mit.edu/