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).
           in_front_of_car(b). pedestrian(b). ...
(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. Abel, D., MacGlashan, J., Littman, M.: Reinforcement learning as a framework for ethical decision making. In: AAAI Workshop: AI, Ethics, and Society (2016)
2. Anderson, M., Anderson, S.: Machine ethics: Creating an ethical intelligent agent. AI Magazine **28**, 15–26 (2007)
3. G.J. Nalepa, M. van Otterlo, S.B., Atzmueller, M.: From context mediation to declarative values and explainability. In: Proceedings of the IJCAI 2018 Workshop on Explainable Artificial Intelligence (XAI). pp. 109–113 (2018)
4. Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: Mapping the debate. Big Data & Society **3**(2), 1–21 (2016)
5. van Otterlo, M.: From Algorithmic Black Boxes to Adaptive White Boxes: Declarative Decision-Theoretic Ethical Programs as Codes of Ethics. In: Proc. of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (2018)
6. Taylor, J., Yudkowsky, E., LaVictoire, P., Critch, A.: Alignment for advanced machine learning systems (2017), mIRI (unpublished) https://intelligence.org/2016/07/27/alignment-machine-learning/
7. Van den Broeck, G., Thon, I., Van Otterlo, M., De Raedt, L.: DTProbLog: A decision-theoretic probabilistic prolog. In: Proceedings of AAAI (2010)
8. van Otterlo, M.: Intensional Dynamic Programming: A Rosetta Stone for Structured Dynamic Programming. Journal of Algorithms **64**, 169–191 (2009)
9. van Otterlo, M.: Solving relational and first-order Markov decision processes: A survey. In: Wiering, M., van Otterlo, M. (eds.) Reinforcement Learning: State-of-the-art, chap. 8, pp. 253–292. Springer (2012)
10. van Otterlo, M.: Ethics and the value(s) of artificial intelligence. Nieuw Archief voor Wiskunde **5/19**(3), 206–209 (2018)
11. van Otterlo, M.: Gatekeeping algorithms with human ethical bias: The ethics of algorithms in archives, libraries and society (2018), arXiv:1801.01705
12. Wiering, M., van Otterlo, M. (eds.): Reinforcement Learning: State-of-the-art. Springer (2012)

[1] http://moralmachine.mit.edu/