Machine Ethics through Machine Coaching

Loizos Michael
Open University of Cyprus &
Research Center on Interactive Media,
Smart Systems, and Emerging Technologies
loizos@ouc.ac.cy

(version: May 13, 2020)

Abstract

We posit that an ethical machine can be constructed by training it towards producing reasons for its decisions that are ethically-acceptable to a human coach who undertakes the supervision of the machine. We propose Machine Coaching as a human-machine interaction paradigm under which the aforementioned training can be carried out.

1 Introduction

Modern decision-making machines are evaluated, for the most part, in terms of their black-box behavior only, and are thought to be successful and ready for deployment if they make correct decisions in certain test contexts. Prior to their deployment, and in anticipation of their evaluation for deployability, machines are trained "in the lab", where their decisions have little bearing (and can cause little harm) on the real world. The training itself happens, typically, through some process of learning, where contexts analogous to the ones that the machine will face during its evaluation are provided to the machine, along with their corresponding gold standard decisions, without any further feedback.

Is the standard practice of developing and evaluating decision-making machines still relevant when the decisions of the machines have explicit ethical ramifications? In this position paper we argue that the answer is negative, and that the development and evaluation of ethical decision-making machines should abide by the following four principles:

(P1) Judge not only a machine's behavior, but also its reasoning.

(P2) Train a machine in real-life situations, and "outside the lab".

(P3) Offer supervision feedback in the form of counter-arguments.

(P4) Acknowledge that ethics is, ultimately, a "subjective" matter.

We put forward the human-machine interaction paradigm of Machine Coaching [Michael, 2019] as a vehicle for developing and evaluating ethical machines. According to this paradigm, a machine is trained in an iterative dialectical manner. During each iteration, the machine faces a context, produces a decision, and provides an explanation for that decision. A human coach then evaluates not the decision itself in isolation, but rather whether the provided explanation offers a sufficient justification for reaching that decision; cf. *(P1)*. Contexts are chosen directly from the machine's intended deployment environment, and the machine's decisions are allowed to have a bearing in that environment, so that the human coach's feedback can also take into account the ramifications of the machine's decisions; cf. *(P2)*. If the machine's explanation is not found to be acceptable, then the human coach offers feedback to the machine, not in terms of what the correct decision would have been, but in terms of why the explanation provided by the machine is incorrect or incomplete as a justification; cf. *(P3)*. Ultimately, after a reasonable number of iterations, the machine converges to reason, and hence reach decisions, in a manner that is ethically-acceptable to the human coach; cf. *(P4)*.

In the sequel, we argue in support of the four principles, and for the use of Machine Coaching for Machine Ethics.

2 Justification of Action

Both in our commonsensical understanding of ethical behavior, and in the way that societies have chosen to formalize some of the tenets of ethical behavior through social norms, regulations, or laws, an action cannot be judged independently of any context in which it has happened, nor independently of the reason that the actor had for taking that action. Courts that interpret the laws of a country pay, for instance, special attention to whether there was an intent behind an action, whether the intent itself was aimed to cause harm, and whether the plan to cause harm was conceived on the spur of the moment, or was devised a priori. Thus, a judge will differentiate between a human driver who runs over a pedestrian: unintentionally because they were absent-minded; intentionally because they chose to avoid running over a group a children who were crossing the street; with a malintent because they were trying to scare the pedestrian who was seen littering; or with a planned goal to kill the pedestrian as payback for something that happened in the past.

An analogous approach should be followed when an autonomous car runs over a pedestrian. We cannot simply exclaim "AI is a killer!" on the first sight of the accident and adjourn. Instead, we should be able to proceed as a judge would do in the case of a human driver, and try to understand the reasoning behind the car's action in the particular incident. Current Machine Learning technologies used in autonomous cars (and in most other ML-powered systems) are ill-fit to provide such justifications, and even if some justifications are available under certain conditions, those are mostly geared towards ML-experts, and would not be readily comprehensible by an average person or a judge.

Machine Coaching explicitly supports the evaluation of the machine's reasoning, and whether that reasoning justifies any resulting decision, irrespectively of whether the decision can be, in isolation, deemed to be acceptable or not. In principle, the machine need not even make a definite decision in cases where this is not needed (e.g., if acting as an assistant to a human decision-maker), and can instead provide explanations for mutually incompatible decisions. In such cases, evaluating the decisions no longer makes sense, whereas evaluating whether each explanation sufficiently justifies its associated decision is still a valid consideration. In all cases, the evaluation happens in the particular context on which decisions were made, keeping, thus, the evaluation task cognitively-light for the human coach.

3 Continuous Feedback

In human societies, more now than ever before, life-long learning has become a central tenet in our fast-changing (at least in terms of available information) environment. Learning (especially at the tertiary and professional levels) is increasingly not seen as something that precedes a human's other life obligations, and that is done in specialized safe places (also known as "schools"), but rather as something that happens in situ, at the place of its eventual use.

Machines would, analogously, benefit from adopting a continuous learning approach to cope with changes in the contexts that they may face, or in the acceptability of the decisions and the explanations they may provide (e.g., due to a new law that might make it acceptable to pass a red light when you are not blocking the traffic). Much more than as a sufficient mechanism to cope with a changing environment, however, a continuous learning approach is a necessary mechanism to support the shift on how machines are to be evaluated. Unlike typical Machine Learning where the gold standard decision is independent of the machine and can, thus, be identified and provided upfront, when evaluating a machine's reasoning it is not feasible to provide upfront a reaction to each possible machine-generated explanation.

As anticipated by Machine Coaching, the human coach engages with the machine in a continuous manner, reacting to the machine's decisions and explanations as the machine perceives different contexts. This, in turn, implies that the machine cannot be fully trained "in the lab", but has to face, early on, the hurdles of real life. Such an in situ training of a machine, not unlike the in situ training of young children, raises interesting ethical issues on the real-life ramifications of the decisions of an insufficiently-trained machine. We do not elaborate further on this aspect, but we direct the interested reader to our past thoughts and technical work on this matter [Michael, 2015a; 2015b].

4 Learning vs Coaching

The feedback that a young child experiences after a mischief is, typically, some negative reward that materializes when the offending action is identified by a responsible adult. What does the child learn from such feedback? Does the child learn that doing mischiefs is wrong and should be avoided altogether, or that a mischief should be followed by an

attempt not to get caught, and hence not to have the negative reward materialize? In truth, we cannot control what one might end up learning from such sparse supervision feedback.[1] It is for this reason that nurturing children is not just a matter of applying supervision through feedback of the form "That is wrong!" or through negative reward of the form "No dessert for you tonight!", but rather a matter of making them understand why a certain action is not acceptable.

The same lack of control of what is learned applies also to machines trained through typical Machine Learning. The decision-making behavior of the machine might end up conforming to the feedback it receives, but the reasons justifying that behavior might be arbitrary. Providing as feedback the entire justification that we would find acceptable for the machine to employ for each particular context is as infeasible for machines as one could imagine is for humans.

By contrast, Machine Coaching embraces the Goldilocks principle on the role of the feedback, as being more than a simple label or a reward — as used in typical Machine Learning, which although very cognitively-light for the human, it is computationally-demanding for the machine — but still less than a full-fledged explanation — as used in typical Machine Programming, which although very computationally-light for the machine, it is cognitively-demanding for the human. Striking a balance between these two extremes, feedback in Machine Coaching comes through a counter-argument to the current justification offered by the machine, appealing to the innate ability of humans to challenge arguments raised by others [Mercier and Sperber, 2011; Mercier, 2016], while also facilitating an elaboration tolerant integration of the counter-argument into the machine's knowledge base [McCarthy, 1998]. Furthermore, this feedback process is accompanied by formal guarantees on how efficiently and to what extent the machine converges to a state of offering decisions and associated explanations that are acceptable to its human coach [Michael, 2017; 2019].

5 "Personalized" Ethics

In this position paper we have discussed a human-machine interaction paradigm and an algorithmic protocol through which a machine can be developed that reasons in a manner that is ethically-acceptable to the machine's human coach. In terms of building ethical machines, then, *we would argue that as long as the human coach is supposed to be ethical, under whatever semantics this supposition is interpreted, then the machine should also be taken to be ethical under that same semantics*. We conclude this position paper by touching briefly upon certain issues that relate to this argument.

The premise of the argument that "the human coach is ethical" does not presume to take a stance on the kind of ethics that the human coach could or should have. We understand that every individual has their own personal version and variant of ethical standards — if one adopts a subjective view of ethics — or their own personal extent to which they abide to ethical standards — if one adopts an objective view of ethics. Still, large communities of people exist that acknowledge that their ethical commonalities outweigh their individual ethical differences, and accept someone as being ethical based on these ethical commonalities. Presumably, then, if one wishes to adopt the ethical standards of a certain community for a machine, one would need to identify a human coach that meets those ethical standards themselves. Alternatively, one could take the human coach to be an appropriately interdisciplinary group of people that (irrespectively of their personal ethics) are able to judge whether others (the machine in this case) are ethical. The group's work could be aided by having, to the extent this is deemed beneficial, the machine extract ethical standards through autodidactic learning [Michael, 2008] from written text [Michael, 2009] coming even from the Web [Michael, 2013], with the provision that since such ethical standards might, depending on the source of data, be heavily biased, the machine coaching process will need to follow autodidactic learning to clean up and debug what has been learned.

In terms of the conclusion of the argument that "the machine is ethical", we are effectively stipulating that a machine developed through Machine Coaching approximately replicates its human coach's ethical standards and should, by extension, be considered to be ethical in the same way, or to the same extent, as what its human coach reveals through the interaction with the machine. Would this approximate replication of ethical standards to the machine make, then, the human coach (and them alone) responsible for the ethical violations of their coached machine, once the latter is sufficiently trained by the human coach? We eschew answering this question, but we hypothesize that the answer should not be inconsistent with those of the following questions: Are parents responsible for their grown-up children's ethical violations? Are two twins raised under the same ethical standards, responsible for each other's ethical violations?

[1] In fact, at least for the case of young children, the feedback that they are given by their parents is sometimes explicitly pushing them to learn to avoid getting caught, rather than avoiding doing the wrong thing, as in the case when a parent might say "*Don't let me see you* do that again!".

Acknowledgements

This work was supported by funding from the EU's Horizon 2020 Research and Innovation Programme under grant agreements no. 739578 and no. 823783, and from the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination, and Development. The author would like to thank Laura Schelenz and Karoline Reinhardt for their comments and feedback on a draft version of this paper.

References

[McCarthy, 1998] John McCarthy. Elaboration Tolerance. In *Proc. 4th International Symposium on Logical Formalizations of Commonsense Reasoning*, pages 198–216, 1998.

[Mercier and Sperber, 2011] Hugo Mercier and Dan Sperber. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences*, 34(02):57–74, 2011.

[Mercier, 2016] Hugo Mercier. The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences*, 20(9):689–700, September 2016.

[Michael, 2008] Loizos Michael. *Autodidactic Learning and Reasoning*. Doctoral Dissertation, Harvard University, Cambridge, MA, U.S.A., May 2008.

[Michael, 2009] Loizos Michael. Reading Between the Lines. In *Proc. 21st International Joint Conference on Artificial Intelligence*, pages 1525–1530, Pasadena, CA, U.S.A., 2009.

[Michael, 2013] Loizos Michael. Machines with WebSense. In *Proc. 11th International Symposium on Logical Formalizations of Commonsense Reasoning*, Ayia Napa, Cyprus, 2013.

[Michael, 2015a] Loizos Michael. Introspective Forecasting. In *Proc. 24th International Joint Conference on Artificial Intelligence*, pages 3714–3720, Buenos Aires, Argentina, 2015.

[Michael, 2015b] Loizos Michael. The Disembodied Predictor Stance. *Pattern Recognition Letters*, 64(C):21–29, October 2015. Special Issue 'Philosophical Aspects of Pattern Recognition'.

[Michael, 2017] Loizos Michael. The Advice Taker 2.0. In *Proc. 13th International Symposium on Commonsense Reasoning*, London, England, U.K., 2017.

[Michael, 2019] Loizos Michael. Machine Coaching. In *Proc. IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI @ IJCAI 2019)*, pages 80–86, August 2019.