Approaches of Machine Ethics - Bottom-Up vs. Top-Down

Author: Christopher Mühl
Normative Reasoning and Machine Ethics WS19/20
Supervisor: Prof. Dr. Christoph Benzmüller

March 25, 2020

Nowadays, our everyday life is based on tools, gadgets, machines, and other technologies. Those have different intensities of ethical implications. E.g. a toaster has comparatively low ethical effects, whereas the software of self-driving cars has the potential to cause millions of serious accidents. The current development of artificial intelligence in research and industry promises an increase of technologies whose ethical impacts are high. Moreover, moral machines are conceivable, i.e. AI systems that behave in a moral way, regarding involved parties (especially humans). In order to implement such systems, researchers examined fundamental approaches to implement ethical machines. In this work, we introduce these approaches, compare them and describe each an example of the bottom-up and the top-down approach. Furthermore, we discuss advantages and disadvantages of the fundamental approach categories.

1 Introduction

With the growing research in *artificial intelligence* (AI) another field opened up, namely *AI ethics*. The objective of AI ethics is to minimize harm that is at least partially caused by AI systems. In 2006 the philosopher James Moor identified a hierarchy of four different ethical agencies [1]. Accordingly, the simplest one is called *ethical impact agent*, which means every machine that has any ethical impact. For example a toaster influences humans' eating choices which in turn affect their health, that is why a toaster can be considered an ethical impact agent. The next agency, the *implicit ethical agent*, comprises all ethically behaving machines that are not able to argue about ethics within behavior. In contrast, *explicit ethical agents* are ethically behaving machines that can formalize and argue about ethical behavior. Finally, the most advanced ethical agency

is the *full ethical agent*. Such a machine is able to reason on professional human-level, i.e. to formulate and reasonably justify ethical judgements.

One important branch of AI ethics is the research regarding machines to behave in a moral way, *machine ethics*. Thus, the examination of the latter three ethical agencies is the objective of machine ethics. While the mere possibility of a full ethical agent's existence is questioned, there are approaches to implement both implicit and explicit ethical agents. In this work, we address such approaches: Section 2 gives an overview of different approach categories. In Section 3, we will compare these categories by means of benefits and drawbacks. Afterwards, we will have a look at two examples (see Section 4) and finally conclude this work in Section 5.

2 Categories of Approaches

In this section, we introduce the fundamental categories of approaches to machine ethics according to Wallach, Allen, and Smit [2]. The authors distinguished bottom-up from top-down approaches and assumed the need for a combination of both for realistic applications.

2.1 Bottom-Up

Since machine ethics is based on philosophy and engineering, the approach categories are adopted from these fields. In engineering, the term bottom-up denotes a procedure to approximate a target using a performance measure. On the contrary, bottom-up approaches in a philosophical sence consider normative values as being implicit in behavior instead of resulting from an explicit formulation of a general ethical theory (a complete set of rules). These meanings are merged into the machine ethics sense of a *bottom-up* approach: Creating an accurate account of a machine's understanding of its own and others' morality by means of a performance measure and without a general ethical theory.

2.2 Top-Down

Analog to the definition of bottom-up, the term top-down refers to the decomposition of tasks into simpler tasks in the engineering sense, whereas in philosophy top-down approaches are based on a general ethical theory. Such a theory determines the solution for every specific case and is therefore very abstract. In order to implement an ethical machine via a *top-down* approach, an ethical theory is selected first, then its technical requirements are analyzed and decomposed before the components are implemented.

2.3 Hybrid

When both the bottom-up as well as the top-down approach are combined we speak of a *hybrid* approach. Although this definition is not very precise, the authors of [3] propose

to implement discrete skills by means of specific AI techniques that can be associated with either top-down or bottom-up approaches. Some skills that seem to be necessary at least for human-like robots with moral behavior are the following:

1. Following Deontological Rules: This is the main capability required for ethical agents, although the subsequent skills are arguably in turn essential for this capability. While the authors propose *supervised learning* to achieve it, we will consider different techniques in Section 4.

2. Socialization: Probably important for human's acceptance of robots is its ability to socialize with humans. This capability could be achieved by *reinforcement learning* (RL). It can be thought of as training the robot's interaction directly by human's feedback.

3. Optimizing Multi-Party Interests: This is a purely computational capacity which is why it might probably be the first one to reach a quality superior to humans. The *applied game theory* is a good approach for this skill since it applies the research field that deals with optimizing multi-party interests.

4. Uncertainty Management: Accurate handling of one's environment and the included agents requires good prediction of the future. For all observers (including humans and robots) it holds, that they can not predict the future exactly, due to measurement imprecisions and chance (if the universe is non-deterministic). Therefore, every prediction is only a guess with a specific probability of occurence. *Probabilistic programming* deals with such uncertainties.

5. Modeling Interactions: As discussed in the previous point, predicting the future eases interaction. Another capacity that supports predicting the future is the theoretical modeling of the world and the processes or interactions within. This can be done be *scenario rendering*.

6. Modeling Intent: The idea of both previous points can be extended when considering interaction with humans in particular. The prediction of a human's behavior is done at best by understanding her intent. One approach to this understanding is the *inverse reinforcement learning* (IRL) where in contrary to RL the behavior of another agent is observed and modeled.

These skills are based on further capabilities that are basically measurement techniques to perceive the environment. Examples are computer vision and natural language processing.

3 Comparison of Approach Categories

Both pure approach categories bottom-up and top-down are fundamentally different. In this section we point out inherent properties of each and consider their benefits as well as their drawbacks.

3.1 Properties of Bottom-Up Approaches

The nature of bottom-up approaches is an incremental approximation towards a desirable goal even if that goal is not explicitly defined. Since a full moral behavior is very complex it is very hard to achieve it as a whole in terms of bottom-up approaches. Therefore, Wallach et al. [2] assume the necessity to reach full moral behavior from the composition of simpler behaviors. The possibility to compose multiple behaviors in bottom-up approaches is an important benefit that reduces complexity. Another great advantage of bottom-up approaches is the independence from a general ethical theory. Thus, the problems of formulating and interpreting such a theory need not to be solved. A third major benefit is the incremental improving that enables the monitoring of progress and the application of machines that have not achieved the final goal but are good enough for specific tasks.

Naturally, there are drawbacks as well. One major disadvantage of bottom-up approaches is their non-transparency, meaning that - ignoring measurement errors - the behavior of incrementally trained machines is never completely predictable. In other words, regardless the number of known scenarios, an implicit ethical machine could potentially behave in an unacceptable manner in unknown scenarios. That fact constitutes an ever-present threat. Learning or training always depends on information. This information might be biased (e.g. racist employment decisions) or contain defective data what can result in undesired training outcomes. This drawback can be avoided by carefully controlling input data, but is very difficult. Furthermore, as generally in supervised ML, training data are rare and expensive. This problem could be faced by artificial training environments, where data is cheaply generated. Unfortunately, such environments are not existent in adequate quality today.

Besides the recognized benefits and drawbacks, there are uncertain problems. One of them is the question if morality is able to emerge from assembled components. Hence, it is not known if bottom-up approaches could achieve morality at all. A second problem is the way to verify morality. If it is only tested in a number of scenarios, the problem of non-transparency as discussed above still threatens. If verified by explicit explanation, bottom-up approaches can not be used alone since they do not provide explicitness (at least not until now).

3.2 Properties of Top-Down Approaches

In contrast to bottom-up approaches, top-down approaches inherently provide explicitness. This automatically involves transparency and facilitates verification. Another big benefit over bottom-up approaches is the possibility of fitting an explicit ethical machine to new contexts by adding corresponding rules. This assumes the ethical machine to have achieved morality in different contexts before.

The above mentioned benefits do not come without costs. These costs are inter alia the formulation of adequate universal rules. Such rules have to define accurate behavioral prohibitions, permissions, and obligations for every possible scenario. And thus, these rules are either flexible and hence abstract or detailed and therefore inflexible. The more

precise they are, the larger the set of rules must be. With higher vagueness of the rule system the more degrees of freedom exist in terms of interpretation. This interpretation could diverge from machine to human in the worst case. In any case, the interpretation of or the calculation of moral behavior based on finite rules is potentially very complex and could even be undecidable. Contrary to the bottom-up category, a top-down approached ethical machine can not be applied before it is complete. E.g. it is not possible (or at least very difficult) to use such a machine without a fully implemented logic inference system or incomplete rules while an insufficiently trained bottom-up machine could behave morally enough for some tasks.

Aside from the above discussed advantages and disadvantages, we point out some problems that are partly attributable to other research fields (mainly philosophy and computer science). One obvious problem is the selection of ethical rules since even humans have not definitely decided for one universal ethics. Popular ethics are consequentialist, deontology, and virtue ethics. While consequentialist ethics is probably the most complex one when it comes to computational effort, deontological ethics seems to be the easiest one to implement. But maybe humans would never fully accept machines that are based on deontology. The solution might be the implementation of virtues or a combination of several ethics. Other open problems are the decision of what to do when several rules conflict or when to terminate the calculation for the best (or even an acceptable) action.

4 Examples

Until now, there exist only some approaches to ethical machines. And these are neither completely implemented nor aim for a full ethical machine. Nevertheless, we describe two examples in this section.

4.1 Cooperative Inverse Reinforcement Learning

A bottom-up approach from 2016 that aims at implementing an implicit ethical agent is *cooperative inverse reinforcement learning* (CIRL) [4]. The idea is that a robot observes a human and thereby estimates her intent. That way, the robot can learn moral behavior and learn what it should do and what not. The authors claim, that their approach enforces active learning and teaching what is proposedly more efficient than passive learning.

Slightly simplified, a CIRL game is a two-player Markov game between a human H and a robot R where both observe the current state s_t and select action a_t^H or rather a_t^R at each timestep t. Thereafter, they receive the joint reward r_t and accordingly update their behaviors π^H and π^R. The behavior here is a function with input an observed state and output a selected action and this function is based on the observation history.

This approach was successfully tested on a 2D discrete grid navigation problem. The authors further mentioned that CIRL is a *partially observed Markov decision process*

(POMDP) which in general are NEXP-hard, but due to specific properties, CIRL has an exponentially lower complexity.

4.2 Logic and Knowledge Engineering Framework and Methodology

Not only an approach to implement a moral machine but also a framework for normative reasoning tools is the *logic and knowledge engineering framework and methodology* (LOGIKEY) [5]. Since it is based on logic and rules it belongs to the top-down category. Further, LOGIKEY constitutes a part of an explicit ethical machine which we will discuss below.

LOGIKEY comprises three layers where the content of each layer is interchangeable. The first layer (L1) contains the logic inference system implemented in modal and norm-based logic. The second layer (L2) contains an ethical theory which can be formulated in standard deontic logic, dyadic deontic logic or input-output logic for example. Finally, the third layer (L3) contains application data and/or logic. In a case study the authors implemented each the general data protection regulation and Gewirth's principle of generic consistency. These case studies where to test whether such specific applications can be formalized and if the whole framework is able to infer adequately which it did when using the right logic for implementation. Between the three layers that include possibly different logics the higher order logic (HOL) is used as meta-logic to translate in between.

Note, that LOGIKEY actually consists of an *ethical reasoner* (L1 and HOL) that calculates valid actions and provides explanations for them, and formalized rules (L2 and L3) that are the basis for the ethical reasoner. To be applied for real-world tasks, LOGIKEY needs a further component that perceives the state of the environment and formalizes it as well as a set of possible actions to communicate these formalizations to the ethical reasoner. Together with such an *AI reasoner*, LOGIKEY would be an explicit ethical agent.

5 Conclusion

This work aims at providing an overview of the different approaches to machine ethics. We therefore introduced the field in Section 1 and defined the fundamental approach categories in Section 2. The main part of this work is the analysis of the advantages and disadvantages of bottom-up and top-down approaches in Section 3. The examples in Section 4 purpose a reader's more concrete understanding of each category rather than a detailed description or examination of the examples.

In summary, the field of machine ethics is still young and implementations are far from usable. Although, there are promising ideas like CIRL [4], LOGIKEY [5], and the fundamental approach categories [2] and important properties were already detected. I personally recommend to approach different skills separately in order to reduce complexity and concur Wallach et al. [2] about the need for hybrid approaches. Moreover,

I emphasize the need for an explicit understanding and articulation of moral action in ethical machines, however I assume the humans' acceptance of ethical machines depends on our capacity to predict their behavior. Hence, virtue ethics that emphasize the cultivation of good character may play a key role in contrast to consequentialist ethics that might be performed in a far more advanced way than we humans are able to achieve.

References

[1] J. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intelligent Systems*, vol. 21, pp. 18–21, Aug. 2006. DOI: 10.1109/MIS.2006.80.

[2] W. Wallach, C. Allen, and I. Smit, "Machine morality: Bottom-up and top-down approaches for modeling human moral faculties," *AI Soc.*, vol. 22, pp. 565–582, Apr. 2008. DOI: 10.1007/s00146-007-0099-0.

[3] E. W. James Uhlmann Emelie Sydney-Smith, *EthicsNet - A community teaching preferred behavior to machines*, https://www.ethicsnet.org/, [Online; accessed 02-March-2020], 2018.

[4] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "Cooperative inverse reinforcement learning," Jun. 2016.

[5] C. Benzmüller, X. Parent, and L. van der Torre, "Designing normative theories of ethical reasoning: Formal framework, methodology, and tool support," Aug. 2019.

[6] L. Lamport, *Latex: A Document Preparation System*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1986, ISBN: 0-201-15790-X.

[7] A. Winfield, K. Michael, J. Pitt, and V. Evers, "Machine ethics: The design and governance of ethical ai and autonomous systems," English (US), *Proceedings of the IEEE*, vol. 107, no. 3, pp. 509–517, Mar. 2019, ISSN: 0018-9219. DOI: 10.1109/JPROC.2019.2900622.