On implementing the three major moral paradigms (Virtue Ethics, Consequentialism, Deontology) in artificial agents

Author: kd120

Based on topic 6 in group 2

1 Introduction

I aim to present a comprehensive analysis of the three classical moral paradigms in the context of building ethical AI, as clearly and as categorically as I could, predominantly from a practical and implementation point of view.

1.1 Background and motives

The first question really is why shall we go through the ordeals of building normative ethics, fruits of human philosophy, into AI agents. It is fitting to briefly discuss this question before plunging into the details of different moral frameworks used in AI, as well as their prospects and perils.

In recent decades, Artificial Narrow Intelligence(ANI) have achieved considerable success like human-level control in solving Atari games (Mnih et al., 2013), DeepMind's AlphaGo beating top human players (Silver et al., 2016) etc., although Artificial General Intelligence(AGI) is still far away and foreign, not to mention the more omnipotent and fictious "Super Intelligence". Yet there can be no doubt that AI systems have been and will be playing an ever-growing part of our life, both public and private.

However, with mounting concerns such as road safety concerns regarding driverless cars, or ethical considerations regarding the moral status of "Killer Robot" type of Autonomous Weapon Systems(AWS) (Maiyane, 2019), researchers and institutes have gradually realised the importance of ethics in AI-enabled systems, and in many cases ethics has in fact become the limiting factor for the extent to which AI products are cleared for interactions with human (Yu et al., 2018). To ensure the safety in AI operations, to restore faith and confidence of AI consumers, to better understand and control the power of AI with all of its uncertainties, and to better promote AI to advance the interest of human society, it is therefore not only of an endowment, but also of a mission, for AI to have a moral concept of right or wrong.

This I recognise is the practical motive behind the effort of making AI ethical, while I do not intend to engage in an age-long philosophical debate of the moral obligations and rights of machines versus those of man, for arguments on that side go far beyond the scope of topic discussed in this work.

1.2 The three great moral paradigms

In constructing AI ethics, there are three distinctive paradigms that can be used to teach an AI agent how to act, in accordance with values within each corresponding framework. They

are **Deontology**, **Aristotelian Virtue Ethics**, and **Consequentialism**/ **Utilitarianism**. To enable later a more bounded and informed analysis, it is vital the core values of each paradigm can be clearly defined here.

1.2.1 Deontology

One of the most prominent figures in western philosophy, Immanuel Kant, argued that certain actions are permanently associated with certain moral values and that those values are absolute (Kant, 1882). In his proposed concept of "Categorical Imperative" in evaluating actions, this would suggest that for one man to perform an inherently bad action such as to kill or to lie it is absolutely unethical (Piloidis, 2020), regardless of the background or context of the situation. Thus the deontological ethics in general, building from the Kantian approach, is essentially based on duty and rules, where the morality of an action is determined only by the intention(goal) of the subject and his action's keeping with his duty. In short, only if the motivation of an action is of the absolute good can it be judged as moral.

1.2.2 Consequentialism

The principle of Consequentialism is simply everything depends on the outcome. Its slogan which can be found among classic utilitarians cried that a cause can only be right if it leads to the "greatest happiness for greatest number" (Sinnott-Armstrong, 2019). In other words, an action is considered moral solely based on the produced consequences, irrespective of the intrinsic nature of that action or the context of situation. The moral value of an action is evaluated based on the value of outcome(utility) produced. Although in classic utilitarianism there are many types of Consequentialisms that differ in means of evaluation and context, in this work it is the general Consequentialism being discussed.

1.2.3 Virtue Ethics

Aristotle wrote, in "Nicomachean Ethics", that the virtue of a man is his character which makes him a good man. And he said that there exists a "phronimos"(a person of wisdom) that others can listen to about how to best live a life and develop virtue. In his mind virtue can only be achieved in a practical sense, from life experience. While not as often discussed as the previous paradigms, Virtue Ethics is a more complicated moral framework that combines some traits of the other two. Founded on Aristotelian Virtue Ethics, this framework is deontological on the onset, where it needs some absolute rules to establish the sense of virtue; but it also does not object to the consequentialist ideas, as gaining and evaluating virtue is a practical process in which consequences of an action must be taken into account (MacDonald & Beck-Dudley, 1994).

Almost coincidentally, one can identify that consequentialism is directly opposed to deontological ethics as moral rightness is decided on future consequences instead of being pre-determined by duty or rules, while Virtue Ethics takes the middle ground of the two. **It is important to bear these fundamental differences in mind for analysing the three frameworks in relation to AI in following sections**.

2 On the implementation of moral frameworks in AI

2.1 Maximising utility with AI

As consequentialism is deeply associated with ideas of maximum common good(the sense of utility), a computationalist might rejoice in knowing the formulation of utility as a measure of fitness of an action is quantifiable and thus computable. Indeed, such concept is no stranger to many AI and decision-making algorithms that construct and solve essentially an optimisation with a utility function being the objective. One very influential AI implementation incorporating consequentialist ideas is Reinforcement Learning(RL), a machine learning doctrine in which agent learns/updates value functions of states and actions by sampling series of actions of highest values/utilities and receiving rewards according to some reward(utility) functions, via agent-environment interaction (Sutton & Barto, 2018). Notice how in RL the reward signals generated by reward functions in response to actions are the only external feedback agent receives about goodness of those actions, emblematic to consequentialism.

2.1.1 The good

Immediately one can identify a major strength of consequentialist AI is its compatibility with computationalism thus ease of algorithmic formulation and implementation. The other obvious advantage I argue, compared with Kantian approach's strict rules, is the good generality/adaptability which means when facing various scenarios agent can act with greater degrees of freedom while maintaining moral consistency, i.e. agent would still know how to act morally according to utility when the context has changed, as opposed to a deontological agent might suddenly behave unethically because the rules are not designed with that particular context in mind.

2.1.2 The bad

For an AI agent especially, the concept of transparency/explainability is of paramount importance, for sake of safety and public trust. I argue that for a utilitarian agent like RL agent the gravest drawback is the lack of adequate explainability to its own actions. I am not talking about explainability in the sense of mathematical derivation or proof to analyse an action, although for RL agent a sound mathematical relation between all the environment and action variables is hard if not impossible to draw. What I am talking about is merely qualitative explainability of the motivation for choosing that action. Because for such consequentialist AI all it knows and cares about is the action's results, motivation and reasoning would be hidden under countless aggregation of hard truths of environmental physics and implicit inference of data samples, becoming a forever unseen truth.

2.2 Deontological AI, the bad and good

The duty/rule-centric nature of deontological framework shares striking similarities with core principles of early attempts of symbolic AI like the "good old fashioned" AI(GOFAI), where programmers would establish a series of rules coded in an algorithm for machines to execute (Gamez, 2020). Such symbolic approach to AI is deontological in nature in the sense that for an symbolic AI agent it can only have a set of predefined actions to take (and an action can

only be taken if the preconditions for that action have been satisfied), the task is accomplished when the goal state(intention) is reached by the agent. It means regardless of whichever context the agent is in, any action that violates the conditions(rules) is forbidden, and only if the goal state is reached can the agent be deemed successful(moral). This also implicitly suggests that if morality is to be instilled in such symbolic way, then programmers must design a set of rules/conditions that are perfectly moral and complete, since these rules will dictate which actions are right for the agent to execute at all time. To formulate rules for every situation is near impossible.

One interesting example of deontological AI is the Isaac Asimov Laws of robotics(Clarke, 1994), where Asimov tried to address machine ethics by the introduction of three absolute laws but later encountered a series of logical and moral dilemmas. The problem with this rule-based theory is when an AI agent is following one law it may have already infringed others, i.e. when an AI tries to save a human it might do harm to other people in the process, or in case of real-world systems, Asimov rules can be in direct conflict with the system purpose, say a military drone armed with missiles carrying out an attack violates the first law that robot cannot harm human (Kim, 2018). In short, the biggest disadvantage of deontological AI is its impracticality, rooted in its lack of the sense of context, its difficulty of implementation and likely conflicts with real-world applications. Let's face it: if the deontological way of doing AI is so successful then why has machine learning which is a method not by pre-set rules but by discerning underlying data structure become mainstream now?

Admittedly there are still some benefits in using Deontology. Namely it is, I argue, much more explainable than a consequentialist RL-like agent, simply because the "if-then" style logical structure of deontological AI is, with its absolutism nature, a hard objective fact irrespective of changing context. Thus it is more straightforward for people to follow, especially compared with black-box type of machine learning practices like neural networks. The simplicity and openness of deontological ethics also make the model abstract and thus less prone to errors, as some argued increased complexity leads to more unaccounted mistakes by developers (Kerns & Roth, 2019).

2.3 Machine virtue ethics, best of both worlds?

As discussed before Aristotelian Virtue Ethics stresses on virtue as property of character, therefore under virtue ethics framework the focus is not any specific action, but the character of agent itself. Implementation of such agent-centric framework would require agent to autonomously learn what is good action, from observing and imitating a "phronimos". For example, Apprenticeship Learning as a form of Inverse RL attempts to recover the reward(utility) functions via inference of data samples generated by "expert" (Abbeel & Ng, 2004), just like an newbie driver(agent) learning how to drive by watching instructor's demonstration("phronimos").

One advantage of this, compared with standard RL(consequentialist) approach, is that Inverse RL does not require prior knowledge of the reward functions, on the contrary it learns the reward function subsequently the moral value of action. This trait can be especially useful as

explicit reward functions dictating all the trade-offs can be difficult to define. Virtue ethics agent does exploit the consequentialist means of achieving its goal, that is tracing back through RL-style utility maximisation. Yet these two agents differ at their goals: unlike RL agent's goal of merely finding the action with highest returns, virtue ethics agent's goal is learning what the standard for good actions(virtue) entails. This pursuit for true virtue is guided by an example("phronimos"), and selection of such example for agent to observe and learn from is deontological, as it is of the absolute objective truth. That is why I argue before and now, that virtue ethics combines the best of two worlds: absolute good and practical wisdom.

Though some may argue such algorithm may be too optimistic, as frequent critics to Inverse RL suggest the complete reconstruction of reward function might require an incredible amount of precisely sampled "expert" data (Abbeel & Ng, 2004). Another critique I would imagine is technical hardship of incorporating virtue ethics as the model complexity with more unknowns is higher than in traditional RL, although some are hopeful as new machine learning technologies emerge it would become more feasible (Couttolenc, 2019).

2.4 Summary

Table. 1. summarises all above qualitative analysis.

Moral frameworks	Ease of implementation/ formulation	Explanability/ Transparency	Generalisability/ Adaptability	Overall Applicability
Deontology	Low(+1)	High(+3)	Low(+1)	5
Consequentialism	Medium(+2)	Low(+1)	High(+3)	6
Virtue ethics	Low(+1)	High(+3)	High(+3)	7

Table. 1. Analysis of applicability of the three frameworks, higher the better.

3 Conclusion

Let me close with a hypothetical case. Assume there is a room on fire. There are two people crawling outside the room, injured but safe from fire. There is another man inside the room who is healthy but trapped by the fire. An AI is tasked to save people from fire.

A deontological agent rushes into the scene, it picks up the first person appearing in its sight and leaves because the rules/goals are already satisfied. An consequentialist agent saves all of the two outside, when it tries to enter the room to save the third its sensor says the fire is too dangerous and its calculation suggests the risk of going in is too high, thus it leaves the only person threatened by fire alone. Finally a virtual ethics agent comes and it rushes into the room to save the endangered person first because it has seen that being done by the firefighters.

Now the above case may contain exaggeration and it may not even be perfectly sound, but what I wanted to preach is that in humanity there is something called characters like bravery that is beyond numerical computations. Therefore, morally and humanely speaking, of all three, virtual ethics is the best way forward. Though it may not be the most technically feasible now, we shall have every confidence in the future. After all, for AI to be truly ethical it needs to learn what we hold dear.

References

Abbeel, P. & Ng, A.Y. (2004) Apprenticeship learning via inverse reinforcement learning. *In Proceedings of the twenty-first international conference on Machine learning*. 1. Available from: https://doi.org/10.1145/1015330.1015430 [Accessed: 14th February 2021].

Aristotle, Ross, W. D. & Brown, L. (2009) *The Nicomachean ethics*. Oxford: Oxford University Press. Available from: http://classics.mit.edu/Aristotle/nicomachaen.html [Accessed: 12nd February 2021].

Clarke, R. (1994) Asimov's laws of robotics: Implications for information technology 2. *Computer*. 27 (1). 57-66. Available from: doi: 10.1109/2.248881.

Couttolenc, O. (2019) Approaches to Deploying a Safe Artificial Moral Agent. *montrealethics.ai.* Weblog. Available from: https://montrealethics.ai/approaches-to-deploying-a-safe-artificial-moral-agent/ [Accessed: 15th February 2021].

Gamez, P., Shank, D.B., Arnold, C. et al. (2020) Artificial virtue: the machine question and perceptions of moral character in artificial moral agents. *AI & Soc.* 35, 795–809. Available from: https://doi.org/10.1007/s00146-020-00977-1 [Accessed: 13rd February 2021].

Kant, I. (1882). *The Metaphysics of Ethics.* Edinburg, Scotland, Stanford University Library.

Kerns, M., & Roth, A. (2019) *The Ethical Algorithm: The Science of Socially Aware Algorithm*

Design. New York, Oxford University Press.

Kim, M. (2018) Deontological Ethics. *AI STRATEGY & POLICY.* Weblog. Available from: https://aistrategyblog.com/category/deontological-ethics/ [Accessed: 13rd February 2021].

Macdonald, J.E., Beck-Dudley, C.L. (1994) Are deontology and teleology mutually exclusive?. *J Bus Ethics.* 13, 615–623. Available from: https://doi.org/10.1007/BF00871809 [Accessed: 13rd February 2021].

Maiyane, K. (2019) Ethics of artificial intelligence: virtue ethics as a solution to artificial moral reasoning in the context of lethal autonomous weapon systems. Available from: http://ceur-ws.org/Vol-2540/FAIR2019_paper_22.pdf [Accessed: 11st February 2021].

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M. (2013) Playing Atari with deep reinforcement learning. To be published in *Machine Learning. arXiv.* [preprint] Available from: https://arxiv.org/abs/1312.5602 [Accessed: 11st February 2021].

Piloidis, L. (2020) Ethics in Artificial Intelligence: How Relativism is Still Relevant. Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:sh:diva-41760 [Accessed 12nd February 2021].

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G. & Schrittwieser, J. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature*. 529, 484-489. Available from: https://doi.org/10.1038/nature16961 [Accessed: 11st February 2021].

Sinnott-Armstrong, W. (2019) Consequentialism. *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition).* Available from: https://plato.stanford.edu/archives/sum2019/entries/consequentialism/ [Accessed: 12nd February 2021].

Sutton, R.S. & Barto, A.G. (2018) *Reinforcement Learning An Introduction.2nd ed*. Available from: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf [Accessed: 13rd February 2021].

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018) Building Ethics into Artificial Intelligence. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. 5527-5533. Available from: https://www.ijcai.org/Proceedings/2018/0779.pdf [Accessed: 11st February 2021].