From Artificial Intelligence to Artificial Morality

Moral Thinking **Aditya K. Rao**

PHI477A **190063**

Abstract

We look at the problem to expand from the notion of Artificial Intelligence to Artificial Morality and try to formalise what a deontologist and consequentialist approach towards creating Artificial Morality is like. We look through the examples of human morality and how recreating them for artificial machines will lead. We discuss the involvement of consciousness in this moral landscape. Further, through the example of self-driving cars, we take a look at the interaction between humans and artificial intelligence and expand the same to discuss if the singularity of human and artificial mind would be of any good consequence.

Introduction

The problem of decision-making in self-driving cars has been a crucial matter of debate in the circles of technological giants. In the unfortunate event of choosing between the safety of a passenger and pedestrian, what is the fundamental logic applied by artificial “intelligence”? One is then begged to ask if morality can even be coded into decision-making objects (including the brain). Is there any empirical basis for morality? Also, in an accident/death, who should be deemed responsible? The team that programmed the “intelligence” or the artificial intelligence itself? We don’t concern ourselves with the legal issues and ownership, thereby holding the owner responsible. Instead, a more nuanced question of the source of morality. Intelligence is the ability to acquire knowledge and apply it. The contemporary research to develop intelligence is the quest to replicate empirical data-processing of brains. We train computers in problems where solutions exist, are known and can be deterministically represented. The moral landscape is, however, rigged with unknown and indeterministic problems that would be impossible to code.

Human morality should necessarily be looked upon from the perspective of intent, unlike consequentialism (and utilitarianism), where collective good supersedes intent. A good consequence can only be moral if the intent is good. A wrong intent that leads to good

consequences is reduced to a matter of chance which cannot be modelled into a machine. Most algorithms initially start on randomised chance, and re-inforcing that through morality would make for a disastrous moral machine. We must include aspects of intent while training our machines.

If we look at the moral landscape of machines, the naive approach is to come up with a list of absolute moral guidelines for AI (quite similar to what we get from religion) and let decision-making be entirely based on this. Leaving aside the mammoth task of coming up with such a list of Commandments when most of these tools will be used by a very diverse section of moral groups, we have a more complex problem in implementation. Human consciousness derives other moral codes from the existing set. Should an AI be allowed that, and will it be able to control it from going detrimental to humans?

Through the above set of examples, we understand that if a moral AI is to be trained, it will have to be along the lines of the human mind: simply for the lack of any other known conscious and moral source and since a concrete understanding of human morality is known, it will help to counter any undesired effects of artificial morality. The paper discusses two approaches to construct artificial morality: deontological and consequentialist. The paper tries to understand to what extent human morality should be coded that would render a safe application for users. We further introspect if this artificial morality can be used to solve our moral dilemmas and, in turn, if the singularity of human and machine consciousness is a desirable outcome.

Morality: A Result of Codes and Conducts

The most common answer to the source of morality is a construct of codes and conducts. For instance, religious teachings, constitutional guidelines and legal ethics form the basis of people's moral law that helps them navigate daily life. People succumb to directives from these authoritative institutes instead of relying on their subjective experience. On a practical level, one does not have a rigorous interaction with self to ponder over the moral implications of their daily actions. They see the unfeasibility of such action that would act counter-productive (if not contradictory) to perform actions necessary for their sustenance. They subscribe to such moral authority and remain constrained to it. For instance, a lawyer will find an ethical justification in

the legal code of conduct to defend a repeated criminal. The justification cited are innocent until proven guilty; right to representation in courts, to name a few. Similar arguments are given to many religious practices while discussing ethical concerns regarding entry into places of worship for women, female genitalia mutilation (FGM), and restriction of condoms among the Christian community in Africa; that they are a result of moral teachings of their religion. Even constitutional guidelines sometimes come under direct conflict with social reality. The Lt. Governor of Delhi overturning decisions of a democratically elected State Government of Delhi citing constitutional supremacy is a classic example of codes of conduct being corrupted for personal gains.

To construct a set of codes and conducts applicable to large groups of entities (humans, animals, robots), it is necessary to have a fail-safe mechanism inherently built into the system of ethics. The fail-safe mechanism helps in invoking corrective actions if the need arises. The constitution has a system of checks and balances where each organ of the government (legislature, bureaucracy and judiciary) has the power to question each other and execute course corrective actions. The efficient functioning of such checks and balances systems is critical to laying a solid moral ground for any society. The lack of such institutions in religion and their dependency on texts and scriptures pre-dating the formation of global communities (interaction of people from various parts of the world for trade, communion or war) has given it an unattractive orthodoxy representation in modern times.

Coming to the problem of artificial morality, if we were to design an artificial moral agent based upon a finite set of codes and conducts (Asimov's Laws of Robotics), we would have to construct a system of checks and balances that the artificial moral agent can apply to itself. It is essential to note the self-implicative nature of this. It is not externally applied to the concerned entities. The State acts on itself through constitutional measures at times of crisis in constitutional ethics and not God interfering to save us. Similarly, an artificial moral agent must have self-correcting directives which are executed independent of human intervention.
Moreover, the problematic notion is that morality also changes with time. This change may be in interpreting moral codes for contemporary realities or a complete revision of the code. The computational efficiency makes time run faster in machine space. Simply speaking, a machine

evolves much more quickly than humans. It is more efficient in storing learned knowledge. Thus, the trait of questioning one's morality from time to time (which has been a common practice over human history) is critical but would seem a conflicting command to a machine.

Morality: Subjective Experience

In the last section, the kind of morality discussed was deontological at best. But modern research in artificial intelligence can be best regarded as a consequentialist approach. All research in AI and deep learning ultimately boils down to the optimisation of a cost function. Whether to maximise or minimise the cost function depends on its nature. An AI built to monitor defence strategy would minimise risk to oneself and maximise impact on the enemy. If we try to construct an artificial moral agent based upon consequential ethics, what will be the kind of consequence that we will be interested in? Perhaps emotions like happiness and kindness or pleasure and pain or guilt and honour. Second order factors like risk, casualty, well-being are a function of these underlying emotions. But the problem is how we accurately represent these emotional states in a machine.

Let us think about this human domain. How do we know that a person is happy? We look at external markers like smiles, elevated heart rates, and facial expressions and decipher that the person is happy. But are these parameters absolute, meaning, if a person has all these external markers, can he be deemed happy? Not really. We will find enough evidence of people suffering from acute depression who show normal behaviour. What we actually do is project these external markers onto the subjective definition that we have of being happy and see the amount of overlap. We get a third-person representation of the experience the other person is having. This projection is possible only when we have a subjective understanding of happiness.

If you are an Arsenal fan, you will understand and relate to the happiness of scoring a goal even if the team is Mohan Bagan FC, an Indian team you have no idea about. But you will not really co-relate to the happiness of solving Navier-Stokes Equation or Reimann Hypothesis (currently unsolved problems). What we best get is a projection of that state, given that we, too, have a subjective understanding of that experience. The closer our experience is to another person, the

more we correlate. Even in that case, one might not have accurately recreated the conscious state the other person is experiencing. One will be called a poor human if he/she takes mere smiling as a sufficient marker of happiness. Again, finite states will always be insufficient to represent first-person experiences accurately. A machine lacking consciousness and subjective experience of such emotions will give an even poorer representation of the same resulting in poor training of a consequentialist ethical agent.

Neuro-scientist and philosopher Sam Harris argues that consciousness will evidently emerge as artificial intelligence progress by virtue of progress just like it did in humans. The indeterministic nature of consciousness is the reason why we cannot predict how artificial moral agents will work with conciousness. We have significantly less clue of how our conciousness works.

Understanding the issue of self-driving cars:

We started with a description of the problem of the self-driving cars and how an AI should react to a time of crisis. Let us examine the problem closely and understand why did the problem arise in the first place. Suppose the traffic rules are sufficient in itself that if all of them are executed by all drivers, we will get no accidents. Humans will still get into accidents because of a differential in biological responses to crises and a differential in the priority of traffic rules. If suddenly one sees a human standing in the middle of the road, some will instinctively prefer to apply breaks while some will take a curve around to prevent colliding (we do not assume people would run over the person citing his fault!). The priority on what one is choosing is completely instinctive but may have deterministic reasons from experience. Also, this instinctive decision is not being relayed to drivers behind and ahead of us. So we are again affected by the biological response of fellow drivers to this instinctive action that determines if an accident will occur or not. If all drivers had the instinct to apply breaks seeing a crisis, no accident would occur. If they had a similar response times, no accident would occur. Thu, the difference arises due to the subjective understanding of traffic rules that causes an accident. Also, that people sometimes do not follow traffic rules.

Again in the problem of self-driving cars, an AI can be trained to strictly follow the traffic rules without any need for ethics to be modelled into it. If we give a machine the layout of roads, it will be able to compute trajectories of all cars such that all reach their destination in the

minimum possible time and there are no colliding paths. The problem arises when they interact with humans or a moral entity whose response cannot be relayed or predicted. Similar to the problem of human-human interaction on the road, two entities with a different subjective understandings of the same objective states will always tend towards a conflict. In our daily life, we resolve these conflicts by interacting and finding a middle path, but this is not possible on roads where decisions are made independently and instinctively for human-human and human-(supposed)machine interaction.

Singularity of Biological and Artificial Mind

Moral codes, whether deontological like the ones we get from authoritative instution or consequential that we build from subjective experience of objective states, can in general be said to be heavily influenced by (if not a direct outcome of) our experience. Extending, it is not just a function of our experience but also our retention of those experience. The mind is an interesting entity that is capable of masking experiences when it requires. We mask of bad experiences of some of our actions and commit it repeatedly. Hume argues that “reason alone can never be a motive to any action of the will” and that we require passion to truly believe in our action. True but reason is a necessary condition if not sufficient condition for morality. The need for passion can be only explained by understanding consciousness. We have perfected memory storage and retention and with the advent of quantum computers massive server will become palm sized. In terms of deductive reasoning, an artificial morality will out play human morality. If consciousness emerges by virtue of progress in artificial intelligence, the slightest of variation in passion/motive for biological morality and artificial morality will lead to catastrophic results without artificial morality deeming itself evil or immoral.

Think of the concern we share about ants. We share none. But we don’t go out in the world destroying the ants merely out of passion. However, when you to construct a building, you clearly deem the existence of ants insignificant. This is the kind of relation we will have with artificial morality if it develops consciousness by virtue of progress. Assisting human to solve ethical dilemas will of least consequence.

Conclusion:

But simple mechanical representation, both deontological and consequentialist approach to make a practical artificial morality will be insufficient. We will just create artificial intelligence where the later restricts itself to finding patterns and solving deterministic problems. As we enter moral landscape, the indeterministic nature of human emotions (happiness, pain, pleasure, guilt, honour) is the biggest roadblock. Even after that without the evolution of consciousness in artificial machines will deem any such representation worthless again due to a lack of passion for Humean morality. Further, a singularity of human and artificial morality will simply be biased toward artificial morality owing higher retention of experience and computative efficiency to reason on those experience.

Reference

1. Misselhorn, C. Artificial Morality. Concepts, Issues and Challenges. *Soc* 55, 161–169 (2018). https://doi.org/10.1007/s12115-018-0229-y
2. Asimov's Laws for Robotics: https://www.scientificamerican.com/article/asimovs-laws-wont-stop-robots-from-harming-humans-so-weve-developed-a-better-solution/
3. Humean Philosophy: https://plato.stanford.edu/entries/hume-moral/#inmo
4. Can we Build AI without loosing control over it?: Sam Harris, TED Talk *https://youtu.be/8nt3edWLgIg*
5. A. F. Winfield, K. Michael, J. Pitt and V. Evers, "Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems [Scanning the Issue]," in Proceedings of the IEEE, vol. 107, no. 3, pp. 509-517, March 2019, doi: 10.1109/JPROC.2019.2900622.
6. Nath, R., Sahu, V. The problem of machine ethics in artificial intelligence. *AI & Soc* 35, 103–111 (2020). https://doi.org/10.1007/s00146-017-0768-6
7. Véliz, C. Moral zombies: why algorithms are not moral agents. *AI & Soc* 36, 487–497 (2021). https://doi.org/10.1007/s00146-021-01189-x
8. AlphaTensors: https://www.deepmind.com/blog/discovering-novel-algorithms-with-alphatensor

9. Fawzi, A., Balog, M., Huang, A. *et al.* Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* 610, 47–53 (2022). https://doi.org/10.1038/s41586-022-05172-4