Exploring moral behavior of thinking machines

Abhi Agarwal

Did morality evolve? Frans de Waal centralizes this idea and furthers it by asking how morality evolved. Opposing de Waal on the idea are a small group of biologists. They argue that morality is a construct or an idea that is unique to human beings, and the purpose of morality is to minimize our animalistic instincts. De Waal gives accounts for the idea that morality has evolved continuously and has evolved from Chimpanzees, Bonobos, and Great Apes. He examines different species of animals and points out that primates close to us in the ladder of evolution display a large number of moral traits. Traits such as peacemaking, empathy, sympathy, and many more. For example, if two male chimpanzees are in a quarrel, the female chimpanzees will often mediate and help them settle their quarrel in order to restore the balance in the community.

The smaller group of biologists consists of credible individuals such as Thomas Huxley and George Williams. De Waal, in his book, coined the term Veneer theory to en-

capsulate their arguments. Veneer theory "assumes that deep down we are not truly moral. It views morality as a cultural overlay, a thin veneer hiding an otherwise selfish and brutish nature" (De Waal, 6). The Veneer theory disregards any connection between morality of humans and the tendencies animals have, and so implies that animals and humans are distinct in this regard. De Waal criticizes this theory in his book primarily for two purposes. The first, the theory lacks proof - de Waal points out that there is no empirical evidence to back the theory. The second, to de Waal it seems unlikely that morality is improved by choice and not through genes. I, personally, side with de Waal. I agree that morality is a product of social evolution - I believe that morality is continuously evolving. Through the proposition of this criticism, de Waal implies that the building blocks of moral agency are apparent within primates (such as Bonobos).

Why is it necessary for some moral behavior to have existed in animals? There is not much of a need of moral behavior when an animal is living alone. However, when animals live within packs or groups then the need for restraint on behavior becomes necessary. One of the truths that is taught to us is that we live in groups because the chance of survival and reproduction are much higher than when living alone. In order for animals to be able to adapt in groups they need to change or mask the primal animalistic nature that they rely on for survival when they live alone. De Waal shows this idea

by providing examples of animals that live in groups. In addition, Edward Wilson furthers this idea in his book 'Journey to the Ants'. Wilson describes ant colonies and their successes as a group. Notably, to assist the queen, female ants give up on reproducing to help raise their brothers and sisters. The female ants reduce the competition there is for mating and foster a sense of cooperation within the colony. These traits and common goals improves the colony's chance to exist for decades. In this regard, ants display the trait of altruism.

De Waal points out that there are aspects of morality that are unique to human beings. He notes that the ability to weigh, reason, and judge two separate moral decisions and choose an outcome is one of them. The other is the ability to be impartial and spectate a situation. In addition, there are also fundamental differences in our society and biologically. We have evolved to a point where we don't necessarily form social groups in order for survival, but for cultural interests, religious interests, etc. We also differentiate in the ability to communicate through an established language. This is an important aspect that helps us in making our moral decisions. The biggest differentiator between human beings and animals, to me, is the idea that we're able to express things we're thinking about in a formal and clear manner. We're able to have a discourse about the disagreements we have, and debate whether those disagreements are valid or not. These

aspects are critical in understanding the evolution of our morality.

De Waal manages to establish and solidify the idea that there was an evolution of morality, and the roots of morality can be seen in primates. To me, thinking about the evolution of morality is intriguing. We're attempting to build machines that can think and potentially machines that will walk among us. Soon we will progress into developing artificially intelligent machines. This raises a question, can morality evolve to being implemented digitally? How could morality evolve to thinking machines as it did from primates to us?

There exists a thought problem in the field of Artificial Intelligence that makes this question worth exploring. Hypothetically, I order an AI to build paper clips. The AI would obey its creator and would begin creating paper clips. Eventually it would turn the entire universe into paper clips. The reasoning behind this is that the AI doesn't have an understanding of limitations like we do. An AI only knows its task and that it has to keep working on that task until it reaches a goal. I believe that our understanding of limitations stands from our morality, and our knowledge of what is right and wrong. How can we extend this moral behavior into thinking machines?

What are thinking machines? What exactly are machines that have the ability to make decisions, and mirror our intelligence? In the Aristotelian framework 'virtue' is moral

responsibility, and represents the bestowment of praise or blame. To qualify for this bestowment of praise or blame must be done voluntarily by the agent. In Nicomachean Ethics, Aristotle writes "virtue is about feelings and actions. These receive praise or blame if they are voluntary, but pardon, sometimes even pity, if they are involuntary" (Aristotle, 30). The first condition for our thinking machine must be that it is able to perform actions voluntarily. Through this condition the thinking machine becomes an agent that is morally praiseworthy. The moral actions of a thinking machine should be indistinguishable from any moral person. The second condition is that thinking machines are able to make intelligent decisions. They use previous knowledge and experiences to make decisions, and weigh the outcome of their actions. The last condition is that thinking machines are able to adapt and learn from their experiences - meaning that their 'system' can be adapted by learning new information. The big assumption here is that there is a motivation for a thinking machine to be moral.

Firstly, what does it mean for morality to evolve beyond human beings? Morality evolved from primates to human beings because there were characteristics that made us unique and different to primates. There exists a similar case between us and thinking machines. In addition, thinking machines are much more different biologically to us than we are different to primates. The word evolution, in this context, doesn't necessarily

mean biological evolution. When we design these thinking machines we've to think about what approaches we will take in designing its consciousness and understanding of morality. In this way, we are evolving our morality. Morality is evolving to suit its new surroundings, and we're adapting our morality depending on the unique nature of these machines. For example, machines can't feel pain, and in designing their morality we have to consider this fact. Thinking machines must acknowledge that they aren't able to cause others pain.

Furthermore, we have established that morality is evolved, and that we're exploring the idea of the evolution of morality beyond us. In order to extend our understanding, we have to understand how we become the moral agents that we are. There are certain traits that are inherent to us, but there are also morals that we learn from our surroundings. Things like taking care of your offspring is hard-wired into us, but is the principle that you can't steal from someone else? In the book 'Braintrust', one of the points in Churchland's thesis is the idea that moral codes are, like language, culture specific, and are learnt as we grow up. She also points out that we have a motivation to learn to be moral. The average man learns and stays moral primarily for success in the future where success could be: monetary success, social success, religious, etc. Here, we making another assumption that the future aims of the thinking machines are like ours. We have

designed these future machines to have aims and motivations like ours of the future.

In addition, there is a big discussion in the field of AI regarding how each thinking machine should be 'born'. There are two distinct routes. The first being that thinking machines should be 'born' as children, and then learn and adapt the same way as human children do. The second being that thinking machines should be 'born' as adults, and have programmed some knowledge about their purpose. Both are extremely flexible, and have their own series of pros and cons. The purpose of being born as an adult is to quickly be able to utilize the skill that the thinking machine is created for. If the thinking machine is created to clean the house, then we don't necessarily want to have to wait 21 years before it would be able to do so. However, moral rules would have to be programmed within the adults in order for them to have an understanding of right and wrong. Here, we also further our definition of thinking machines as we establish the idea that their decision-making must be stochastic.

Since we could be able to program a computer to make certain deliberations, what should the deliberations be? Should we even make deliberations? There are four possible ways that researchers have thought to be valid options. They are: Direct Specification, Domesticity, Indirect Normativity, and Augmentation. These are all options that consider different approaches to programming certain sets of information into the thinking

machines, and consider what their motivation would be if that option was considered. What goals would we program a computer to fulfill? Churchland suggests that morals are not objectives or transcendent, but sometimes to us they feel as though they are. Depending on the approach we take, in order for us to allow computers to follow these morals or program them, we have to make them objectives or clearly defined checks that they do.

Firstly, Direct Specification is directly embedding the right set of motivations into the programming for the thinking machines. The difficulty here is "in determining which rules or values we would wish the AI to be guided by and the difficulties in expressing those rules or values in computer-readable code" (Bostrom, 139). This was made popular by Isaac Asimov through his Three Laws of Robotics. The three being: "A robot may not injure a human being or, through inaction, allow a human being to come to harm", "A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law", "A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws" (Asimov, 27). The trouble here is that you can only define them once. They are also open to interpretation - Bostrom points out that the most optimal choice by the AI would be to put the humans into artificially induced comas thereby keeping them safe. For this to execute well we

would have to think about every case and edge-case, and define each thing carefully. In addition, given this option it's difficult to think about how the thinking machine would do in the counterexample to consequentialism: the trolley problem. Bostrom writes "A small error in either the philosophical account or its translation into code could have catastrophic consequences" (Bostrom, 140).

Secondly, Domesticity is an option that limits the scope of the thinking machines ambitions and actions that it can do. The name behind the option comes from the idea of domesticating wild animals. We have domesticated cats and dogs for generations now, and domestication allows us to train animals to lack motivation to perform actions that might harm their owners. We train them to be happy within the domestic environment they live in and to limit their behavior and, by extension, their actions. They are programmed and created to think about and make decisions on a smaller set. A good example for this is in the book 'Hitchhiker?s Guide to the Galaxy' where they create a supercomputer to formulate an answer to a single question. The actions this supercomputer performs are limited, and it's programmed to be able to freely think, but only to think about this one particular question. Bostrom extends this by suggesting that we're able to "'box' an AI such that the system is unable to escape while simultaneously trying to shape the AI's motivation system such that it would be unwilling to escape even if it

found a way to do so" (Bostrom, 141).

This option becomes particularly interesting when you compare it to the Direct Specification method. Bostrom writes, "it seems extremely difficult to specify how one would want a [thinking machine] to behave in the world in general ? since this would require us to account for all the trade offs in all the situations that could arise ? it might be feasible to specify how a [thinking machine] should behave in one particular situation. We could therefore seek to motivate the system to confine itself to acting on a small scale, within a narrow context, and through a limited set of action modes" (Bostrom, 141).

Thirdly, the basic idea behind Indirect Normativity is "rather than specifying a concrete normative standard directly, we specify a process for deriving a standard" (Bostrom, 141). This follows the ideal observer theory in ethics, and we would guide the thinking machine to mirror an ideal and hyper-rational human. The downside here is that we give the thinking machine a lot of leeway, and we are uncertain of the outcome that could produce. In addition, we're also able to offload the thinking machine to do the cognitive work that we needed to perform to come up with a direct specification. Then use the results from the thinking machine to consider the Direct Specification option.

The difficulty behind Indirect Normativity is that we have to consider the normative ethical theories we could utilize in the initial programming of the thinking machine. I

will consider the three competing views, and then discuss specific theories. The three big competing views are: virtue ethics, deontological ethics, and consequentialism. The difference between these views lies in how they approach moral dilemmas rather than the conclusions they reach. The first is virtue ethics, which focuses on the development of an individual's character and the virtues he/she embodies over singular actions. The second is deontological ethics, which judges the morality of a particular action by how closely he/she adheres to a pre-defined set of rules. The last is consequentialism, which argues that the morality of an action should be judged on the outcome.

The issue with deontological ethics is that it requires there to be a pre-defined set of rules, and it measures the sense of morality by how closely you follow them. This approach is definitely something you could utilize when you have a pre-defined set of rules. From the perspective that we need a theory that we're able to programmatically put into a computer, this is not the most optimal choice. The issue with virtue ethics is that it focuses on the development of the single thinking machine's character. Programmatically this is extremely hard to achieve, as it requires a tremendous amount of self-reflection and experiences in the past to look back at. There needs to exist an idea of inner morality. In addition, there is also a debate over what virtues are and what they are not. This makes it extremely difficult to translate into code. The most optimal option

here is to go with the consequentialist view. The philosophies under the consequentialist view that we will consider are: Utilitarianism, Ethical altruism, Teleological ethics, and Preference utilitarianism. Utilitarianism theory aims to maximize utility where utility can be defined in various ways. Preference utilitarianism aims to maximize utility while considering the interests of the people directly involved. Ethical altruism would guide the thinking machines to take actions that would maximize utility for everyone except the thinking machine. Teleological ethics is a theory that utilizes the end consequences of each action to determine which action to take by pre-determining the goodness/badness in performing each action.

Lastly, Augmentation is very similar to the idea of Indirect Normativity, but the idea is to use the physical human brain and extract its 'motivations'. All the other ideas are to build a thinking machine from scratch. This idea begins by taking an existing system, such as a human brain, and utilize the existing intelligence and rules of morality to automatically translate into a thinking machine. The science, obviously, isn't available yet, but this is a serious consideration. It's much easier to use an existing system that has had several centuries to evolve than to create one from scratch. 'With augmentation, we would at least start with a system that has a familiar and human-like motivations" (Bostrom, 142).

All of these approaches above have something in common. They assume that the thinking machine would be created, and then would start being useful from day one. However, if thinking machines were to be born as children, in the same way we are, then they would be able to mirror the development of a human child. They would be able to go through the same process of understanding what not to do, and what to do. This idea is inspired by the view of virtue ethics. We concentrate on the development of the character.

Which of the options are optimal? Domesticity doesn't allow for as much freedom for a thinking machine as Direct Specification does, but reduces the amount we have to generalize our morality. Direct Specification requires us to come up with a lot upfront. The idea of Augmentation is very hypothetical, and would require science to improve instrumentally before that can happen; Augmentation is an idea that is harder to achieve than the creation of thinking machines. The creation of thinking machines would somewhat be a prerequisite to Augmentation as you're basically creating thinking machines, but with the mind of a human being. However, some researchers truly believe that Augmentation could be the answer; it has the benefit of passing on our biological evolution, and the evolution of our morality. Moreover, Indirect Normativity requires a lot of trials. It would take a lot of man-power and cost in order to get right. Thinking machines

would have to be put into the world and used in society.

The more optimal choice would be a combination of these options. The option of Indirect Normativity applied with the idea of Domesticity would work well. Having thinking machines be domesticated and being able to derive a standard from that would be ideal. This way we're able to create thinking machines that are specialized in their tasks.

Moreover, we've established that we're able to create thinking machines and have them establish rules of morality. Now we have consider factors that contribute to our moral decisions apart from our experiences and our virtues. It is argued that emotions do play a key role in our moral decisions. Aristotle's ideology is that moral education must involve learning to feel the "right emotion to the right degree at the right time" (Aristotle, 1104b13). Aristotle view's all of our emotions as "intrinsically relevant to ethics rather than identifying a privileged class of emotion ... In this perspective the value of emotions to ethics lies not so much in what emotions can contribute to our moral behaviour, as in their nature as components of the good life, without which the very idea of morality would be pointless" (Sousa, 1). Through this perspective, emotion is something that the thinking machine would be lacking. Most researchers believe that emotion is something we will not be able to program into an AI. In addition, for most

individuals there is a religious aspect in making moral decisions. Christian's believe in the 10 Commandments, Hindu's follow the words written in the Gita, and most other religions have their own belief systems. These are topics for separate papers, but for the purposes of this essay most researchers believe them to be positive aspects if they were discovered by the thinking machine. Researchers don't believe that a solution can be found for religion or emotion, and if they were to be implemented then they would mirror the Indirect Normativity approach.

Moreover, it's also interesting to explore other ideas that that are a consequence of thinking machines. An idea worth exploring is the idea of perfect knowledge. When two thinking machines are interacting, unlike humans, if given access are able to directly read each other's 'memories' and thoughts. With this, we're able to create a society of thinking machines that have perfect knowledge of each other's thoughts and actions. Given this hypothetical society, the rules of morality would be wildly different; the premise that people instantly know of each other's wrongful actions would reduce immoral thoughts. Given a system that is utilizing the Domesticity approach, each of the thinking machines would have a fixed set of moral values and it would require a majority of the thinking machines to reverse their moral values for them to perform immoral actions. This could apply in high-sensitive places where it is important to work in harmony, such as bank

vaults or stock exchanges.

In conclusion, (We aren't able to predict how society will evolve for thinking machines or if that will happen).

References

[1] Bostrom, Nick. *Superintelligence: The Coming Machine Intelligence Revolution*. Oxford: Oxford UP, 2013. Print.

[2] Asimov, Isaac. *I, Robot*. New York, NY: Bantam, Spectra, 1950. Print.

[3] De Waal, F. B. M., Stephen Macedo, Josiah Ober, and Robert Wright. *Primates and Philosophers: How Morality Evolved*. Princeton, NJ: Princeton UP, 2006. Print.

[4] Aristotle, and Terence Irwin. *Nicomachean Ethics*. Indianapolis, IN: Hackett Pub., 1999. Print.

[5] Churchland, Patricia Smith. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton, NJ: Princeton UP, 2011. Print.

[6] Sousa, Ronald De. *"Moral Emotions"*. 4:109-126. Ethical Theory and Moral Practice. Print. 12 May 2015.