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 encapsulate 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.

Is morality learned? Does evolution of morality mean we take principles we have learnt of morality and teach/embed that within computers?

Arguments for learnt. AI research in starting off like a child.

Arguments for a goodness within us.

What goals would we program a computer to fulfill?

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. (http://philosophicaldisquisitions.blogspot.com/2014/08/
on-superintelligence-6.html)

Morality and its connection to emotion.

Frameworks for morality in the future. Perfect knowledge of all actions.

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

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