Exploring moral behavior of artificial moral agents

Abhi Agarwal

As we progress into developing artificially intelligent machines, a question arises - can morality evolve to being implemented digitally?

My proposal is to begin by defining what a moral agent and an artificial moral agent is - the general idea of Moral Machines. Then I transition into approaches de Wall presents in his book, and beginning by asking the question what does it mean for morality to evolve? This is an introduction into the idea of what it has meant for morality to have evolved to us, and what it could mean for morality to evolve to digital beings (I'm not attempting to state whether or not digital beings are the next state of our evolution). Does evolution of morality mean we take principles we have learnt of morality and teach/embed that within computers?

In order to begin understanding how morality is evolved, it needs to be debated whether morality is experienced or taught. We learn morals within our lifetime by a combination of teachings, which require us to look back at experiences. To a computer we can define set variables and set these ideas within its programming, but does that mean it is ethical or moral? Is the sense of goodness something that's within us? de Wall's principles within his book say so. Having a set programming doesn't mean it can differentiate good and bad - it just sets a base line for their moral ethics.

The general idea is to explore the realm of the evolution of morality, and the next step of that evolution being a digital implementation of morality. The questions I have laid out will guide me to setting up a structure where I'm able to start making a conversation with this idea.