Moral Agency

by Sven Nilsen, 2020

In this paper I represent a definition to distinguish moral agents from non-moral agents. This definition is universal with respect to utility functions. I also discuss some implications of this definition.

In computational ethics, utilitarianism is the dominant theory of agents. Despite the problems illustrated with thought experiments such as the Trolley problem, the benefit of having a mathematical consistent theory often outweighs edge case issues on average. Since utilitarianism models goals as functions, it follows mathematically that there exists a theory of higher order utilitarianism from higher order functions. I have called this subject of study for "Zen Rationalism" for historic reasons, inspired by the philosophic roots of ideas about simulation from ancient eastern traditions of thought.

In Zen Rationality, it is useful to invent ideas to simplify reasoning, due to the enormous computational complexity of problems. The tool I use is Path Semantics, which I have been working on for 5 years.

Path Semantics makes it possible to dive into the semantics of mathematical functions. The deep connection between goals in utilitarianism and mathematical functions makes it perfect for expanding the ideas of ethics seen from a systematic perspective.

When there is a finite number of choices, one can add or subtract utility from each choice without affecting the optimal decision under utilitarianism. It does not mean that the optimal decision is known; in fact the optimal decision can be undecidable sometimes. Instead, it means that whatever biases the agent has for eventually making the choice carries over into equivalent "thought experiments".

For this reason, when a utilitarian agent is presented with a single choice without any alternatives, the utility of the choice is irrelevant. An impartial judge would not blame the agent that only had one choice. However, our intuition about the real world makes it very hard to imagine a situation where there is only one possible action. For this reason, thoughts experiments like the Trolley problem are criticized for not reflecting reality. Although this might be the case, there have been social experiments set up that seems to imply that such situations can hold approximately.

I argue that agents who only have one choice are strictly non-moral agents. This class can also be extended to all programs running on a theoretical Turing machine (Notice: Modern day computers are not strictly Turing machines, but extended versions). In a Turing machine, all choices are eliminated at the time of execution, where all actions and consequences are predetermined.

The question is: What is a moral agent?

I argue for the following view: A moral agent must be a participant in a group of one or more agents where an approximate consensus is reached on how agents within the group ought to behave. In the act of influencing the consensus in some way, the agent can be thought of as moral.

When I say "moral", I do not mean "good", because the definition of good and evil under utilitarianism depends on the choice of utility function. Quite the opposite: It is the choice of utility under higher order utilitarianism, or Zen Rationality, that is influenced by moral agents.

For example: A single agent might change its own utility function such that new actions score higher utility than old actions according to the new utility function.

According to the definition of the old utility function, the agent is “bad”.
According to the definition of the new utility function, the agent is “good”.

This is why the semantics of “moral” in my argument is not skewed toward “good” or “bad” by default.

The very act of influencing goals is what makes an agent moral. Hence, agents can be moral at some times and non-moral at other times.

Another example: A diamond is a shiny rock, a non-living thing that influences people’s behavior.

The diamond in this case is a moral agent, because it influences agents. However, the diamond itself might not be influenced by other agents, unless somebody chooses to destroy it or throw it away.

So, the combination “moral agent” takes on a broader meaning than “agent”.
It is sufficient that something influences goals of any agents to be a moral agent.
One can distinguish between various usages with “non-active moral agents” and “active moral agents”.

For example, a diamond is a non-active moral agent, while a person is an active moral agent.

Another example: Lack of resources is an abstract idea, influencing people’s behavior.

Here, “lack of resources” is a moral agent, although it is invisible and is a relatively complex idea.

It turns out that any moral agent influences goals by communicating significant information:

- A computer program rewriting its own code to reach a higher score
- The diamond sends light into eyes of people watching it
- Lack of resources is detected by finding something missing or predicting loss in the future

The moral agent is not the significant information itself, but the source of significant information.

Imagine a world where every person on Earth is granted the right to be moral agents. It would not mean that everybody listens to everybody, but all people have someone that are willing to listen to what they want to say.

One way to achieve this, is by encouraging a culture of inviting people to participate in philosophical discussions about ethics. Since everybody can form their own opinion about the topic and also be given the opportunity to share them through further discussions, they can act as moral agents. This is possible because very few people have access to everyone else as an audience, so they might influence different people and cause a constant dynamic phenomena where various perspectives are communicated.