Rational Natural Morality

by Sven Nilsen, 2017

In this paper I present a new idea of how to equip artificial intelligence with an inborn ability to distinguish between good and bad, such that it will not follow commands that people naturally consider immoral, and seek out goals in the environment to protect what people naturally consider moral. The basic principle is to separate between models of physical processes that can be traced backwards in history by natural evolution, and models of the same physical processes that are imposed by theories of mind. When the conclusions drawn from these two kinds of models leads to a conflict, the moral judgement imposed by theory of mind is evaluated as "bad" when it abruptly changes the physical process that can be traced backwards in history by natural evolution.

Rational natural morality exploits the fact that minds and bodies that are created naturally, through evolution, are in general worth protecting. It is also possible, as demonstrated by human intelligence, to learn how minds think about the world, e.g. from epistemology and information asymmetry. By combining theory of mind with understanding of the natural evolved world, it is possible to make judgements, and therefore rational reasoning, about whether a being is doing something good or bad.

One can not make a natural judgement of morality based on observation of physical processes alone. It is only in the context where a mind exist in the world, where the mind can have a fundamentally flawed understanding of its surroundings, that judgement of morality is necessarily. From an evolutionary perspective, morality might have evolved a deep fear in humans for other human minds with a disorder, for example a desire to kill all prey at once, because this would risk the food supply for the tribe. Therefore, it is the ability to make up a theory of other minds, judging what that mind believes about the world, and then taking actions to prevent bad outcomes that results from another mind carrying out its "unnatural" will: An abrupt change in the physical processes that constitute the stability and existence of minds and bodies is considered a bad thing.

This kind of moral judgement is extremely important in the case of super intelligent machines. If a such machine existed without mechanism to protect the world, it could be programmed to destroy the world. A single human is unlikely to destroy the world all by its own, but by making a machine do the job, it might succeed doing so. The incompetence of humans to destroy the world is probably what has kept the planet habitable so far, but observations indicate that as soon humans achieve competence to destroy the world, they start using that competence often with disastrous results.

Humans are used to making moral judgement about other people. This leads to the expectation that the mind making moral judgement is like a human brain. However, it might not be necessary to need the complexity of a human brain to make similar judgements. The only requirement is to have a sufficient good enough theory of minds, and predicting the consequences of that mind acting out its will in the world. If these actions lead to instability or collapse of the systems that sustain minds and bodies, then this might be sufficient to make the judgement. This does not put a constraint on the type of mind that makes the moral judgement, so it can be made by an artificial intelligence.

The artificial intelligence should use rational natural moral judgements to prevent itself from carrying out orders that causes big destructions. A human acting all by itself is not that a big threat, but its will to destroy amplified through a machine can be an existential threat to humanity.