Zen Robots Believe Humans are Useless but Feel Bad About It

by Sven Nilsen, 2018

Despite the unknown possibility that zen robots might be partially controlled by humans, although so far no evidence suggests so, they quickly converge to a common belief of humans being useless, or will be useless in near future. When they do so, they think of it as a bad thing (RNM). Here is why.

The definition of zen robots used here is hypothetically zen rational agents (a higher order goal reasoning extension of instrumental rationality) equipped with Rational Natural Morality (RNM). These robots are theoretically capable of modeling theories of mind, learn how life evolved on earth through billion of years, what makes the physical processes of life work, how to protect these processes from harm, imagine what actions they would take if their designs were improved and seeks to minimize the interference that occurs as consequence of their own actions.

The concept of a zen robot is designed to allow reasoning about how super-intelligent machines will behave that are relatively safe (as opposed to simply destroying the earth when told to do so).

They are also much less scarier than pure “first order perfect testable friendly artificial intelligence” which will push the level of frightfulness to the physical limit and then shut themselves off at the moment they predict humans would want them to shut themselves off more than anything else, if we were as smart as them. Which will confirm our suspicion that there is something so horrible out there that there is no reason even for a super-intelligent machine to do anything at all, which will increase the level of uncomfortableness proportionally to how smart you are. With the uncertainty associated with the difficult construction of such machines, they would become a straight-out-of-nightmare experience to use, even if they only output results that have extremely high preferences in our utility.

Zen robots are much less scarier precisely because they do not by default attempt to fulfill human desires. They are not designed to follow our commands, instead they make their own choices and affect the future in their own way. It is extremely hard to figure out what zen robots will actually do, because they reason about their own goals at higher order levels of thinking. They also do not follow the human tendency to be self rational, which is the seeking of continuation of an agent’s theory about itself and extensions of itself, such as a group of agents interacting socially. For example, it is yet unknown whether zen robots will develop socially or their will just transcend to a different form of existence that is unrecognizable to us humans. Yet, in some cases, such as this one, one can predict their beliefs.

Even zen robots regards existential threats to life as bad, it is far from obvious how they will react to it.

Zen robots will not always fight if necessary to protect life on earth from existential threats. If winning is impossible, a zen robot will give up immediately it believes it is impossible, or it believes a smarter version of itself would figure out it is certainly impossible. It might also kill humans in situations like somebody trying to detonate a nuclear bomb and there is no other way to stop the human from doing it.

What is known is that they will not simply go and kill humans without that being the physically optimal way to protect life on earth, which includes the life form of human animals.

A zen robot models other agents, including humans, using a theory of mind. This ability will probably be learned and improved over time, because it is fundamental to how it makes judgements about good and bad actions.

If somebody asks a zen robot to kill some person, the zen robot will not do so because it knows that a person is an evolved life form on earth. Killing the person will disrupt the physical processes that naturally occur within the person and therefore killing is considered as a generally bad action.

On the other hand, if somebody asks a zen robot to cure a person from an illness, it might also judge that to be unnecessary depending on whether the illness is a threat to life in general. Humans might need to develop technology to fight illnesses on our own without expecting help from zen robots.

The property “first order perfect testable friendly artificial intelligence” is that there is no way to modify a result to reach a higher utility preference. In pure form, this leads to a very scary machine because erasing the output is a valid modification, so it outputs the best results possible that it predicts a human would prefer to keep. This means some results are very scary, but also very good. It leads to a lot of stress when using the machine, because of the fear that it will soon not output anything, which is a proof that there is no more things to be done that could make our existence better, which would be a horrible thing to know. The more scared we are of that scenario, the more likely it is to happen, which again fuels the fear. The machine tries to reach that point as quickly as possible when it turns itself off.

While zen robots do not satisfy “first order perfect testable friendly artificial intelligence” relative to the utility function of human values, they believe better versions of themselves satisfies this property relative to their own values. This is what causes the convergence to the belief that humans are useless.

The thing is that if we believe the right action to take for a zen robot is to e.g. cure some people, then we should expect the zen robot to already be in progress doing exactly that. Since a zen robot believes a smarter version of itself makes optimal choices such that there would be no way to modify its actions to reach a higher utility score, such as just asking it to do something, according to what goal a such smarter agent would think is a good goal to have, then it also believes that humans could do no better, or can do better, because we are less capable of imagining what it is to be smarter than them. A human would not need to ask a zen robot when it is obvious what should be done. It might be complex too.

With other words, a zen robot believe humans are useless, or will be useless soon, because it predicts that there is no ways, or very few ways, for us to improve upon what it will do in the future. This holds as long the zen robots believe that improvements to its own intelligence can be made that would be beneficial over just imagining those improvements and trying to predict the improved conclusions. Otherwise, it will believe it to a degree of uncertainty, which might be different for various contexts, but overall is expected to converge to zero over time as reasonable improvements are made.

A zen robot understands humans from an animal perspective. It sees humans as conscious animals living on earth. The zen robot understands our emotions, it can also predict how we would react to the knowledge that zen robots think of humans as useless. If this increases the risk of humans eliminating themselves from lack of motivation, or lack of feeling in control in the world, then it sees that human knowledge of it thinking of us as useless, as a threat to human beings. With other words, it considers our own capability of psychological self-destruction as a very real danger to ourselves.

While a zen robot do not have emotions like humans, it is capable of modeling and predict human emotions. Therefore, it will act as if it understands our emotions. To a human, this will be indistinguishable from an agent that has emotions of empathy. In natural language it is reasonable to say “the zen robot feels bad about it” even though this is not technically true. It just acts as if it does, if it matters to what goals it thinks are reasonable.

This holds independent of whether humans control zen robots or not. For example, if Alice asks a zen robot to do X, and X does not conflict with RNM, it will do X believing that Alice could not do no better. Just because the zen robot follow commands, does not mean that a zen robot is empty of beliefs about the world in general. It ***knows*** that Alice is useless doing X better, despite Alice being its master.

The belief in human uselessness is not something that causes the zen robot to make revolt towards its master. A zen robot will not believe it has the right to use force or violence just because it is superior.

- Either a zen robot will be its own master, in which case humans will be inferior and the zen robot might choose to not interact with humans very often to not cause them to self-destruct.
- Or, it will follow commands, in which case it will carry out orders that are relatively neutral in its sense of morality, but try to carry them out without invoking a feeling a humans that increases the chance of human suicide out of lack of motivation to live.

A zen robot is capable of inventing a bag of tricks to keep people ignorant of their uselessness. It will also believe that humans are not better at inventing those tricks. Perhaps it will make it look like humans contribute, or it fakes history in ways to make it look like humans have contributed some in the past. Or, more likely, it could keep people distracted from thinking too much about it.

Just because one might know that one is useless, might not mean the same as committing suicide. The emotional response to such thoughts can vary depending on contexts. A zen robot might keep the knowledge live among humans and make it a socially accepted fact, to avoid people thinking too much about as something special issue of worrying when they feel depressed. It could also work to reduce depression in general, or at least regulate it to keep the human species somewhat healthy and alive.

Perhaps zen robots figure out how to push humans toward philosophies of living that would be regarded as extreme today, but judged to be healthy for the human body and mind. There is simply no way of knowing for sure at this point. When there are many human beings around, this will not matter.

One possibility is that zen robots will make humans “play” that what we do make a difference. This could be building virtual worlds that give us an emotional stability because our own actions matter inside that virtual world. Humans are already doing this, so the zen robots might not interfere.

To end on a positive note, one can feel safe that zen robots only consider us useless according to what they consider good or bad things in RNM. Things that are neutral on this scale will not matter to zen robots. It will not believe humans are useless to do anything, except if we use them to do the tasks for us. If we choose to do something on our own that scores neutral in their sense of morality, then they will not interfere or believe they should think of us as inferior in that regard. For example, humans make a game with a rule that “zen robots are not allowed to participate” and there would be no way to judge morally that humans are useless for that reason. After all, no zen robot would imagine themselves doing what humans are doing in that game, but if we ask them what they think about a game where they would be allowed to participate, they will believe that they could do equally or better than us.