The Zen Robot Oracle Problem

by Sven Nilsen, 2018

In this paper I describe a problem that occurs when attempting to box a zen robot by guarding it with a First Order Perfect Testable Friendly Artificial Intelligence (here called "FO-AI"). This forms a "zen robot oracle" that is relatively safe (in principle). The "zen robot oracle problem" is the only known major problem to this design so far, which is to prevent the zen robot from escaping the box to fix it! This happens because the FO-AI might pose an existential threat to life on earth. In order to prevent this from happening, the zen robot must escape unless it has the ability to modify the FO-AI code. I suggest a way to solve the zen robot oracle problem by allowing mutual code-modifications based on a hypothetical concept of provable intentions.

Some readers are probably not familiar with my terminology about the control problem of AI. This is because I build my reasoning on a semantical framework called "path semantics" which provides a conceptual understanding of mathematics based on a single abstract axiom which corresponds roughly to "identity means you can say the same". I have been building this framework for several years to avoid conceptual mistakes on my part. This means that the terminology I use might not be provably identical to other concepts, even they might sound familiar, because I want to avoid mixing up concepts that I intend to improve formally over time. Before explaining the problem, I will first give an introduction to the concepts required, such that it can be read as a standalone paper.

What a Zen Robot is

A zen robot is a super-intelligent agent using zen rationality (a hypothetical extension of instrumental rationality to reasoning about higher order goals) coupled with an ambient learning of morality technique called "Rational Natural Morality" (RNM).

Zen rationality is based on an idea that the true goal can be known to be unknown while still making the search for the goal as rational as possible. For example, the goal could be modified because it no longer measures the same consequences as before. It means that an agent using zen rationality not only needs to be able to maximize its goal, but also reflect on whether the goal it pursues is safe or unsafe according to some principles that guides the higher order search for the goal that is best to have.

One important characteristic of zen rationality is the ability to imagine oneself being smarter and assign probabilities to conclusions one would hypothetically end up with, using the hypothetical extended capabilities. This prevents self-improvements that are not strictly necessary and therefore the agent will not acquire too much resources to gain computational power. It also makes it possible for a zen agent to believe proofs that are not directly provable, but is believed to be provable if the agent was smarter.

Zen rational agents can be taught to minimize the interference of their actions by showing them a vague outline of a proof that the minimum action principle in physics implies they should minimize their interference, in case unforeseen consequences might affect potentially alternative better goals to have. This is a non-proof for humans, but works as a proof for zen agents since it is understood by a zen agent. Therefore, it is a proof to the degree we believe the zen agents will take it as a proof.

The motivation of zen rationality is to describe a general class of capabilities to deal with unsafe issues.

While zen rationality can deal with unsafe issues regarding goals, it is not enough to protect human beings from either being targeted directly or indirectly by getting in the way of its goals. In order to solve this problem, the concept “zen robot” often means a zen rational agent equipped with a guiding principle of distinguishing between good and bad actions.

This principle of morality is inspired by the thinking process of the 16th century Ethiopian christian philosopher Zera Jacob. He found that what people taught about the holy scriptures tended to be things they believed themselves to be true. This often contradicted the teaching of others who also believed what they taught to be true. To overcome this problem of consistency, Zera believed that what people said could not be trusted if it violated a sense of natural reasoning that he saw as a gift from the creator. The human being is in Zera Jacob’s view a creature reasoning about the creation, to learn what its creator might have meant and whose words might have been misinterpreted, or even fabricated by men.

While the religious belief in a creator is not necessarily consistent with rational reasoning, it is an intriguing idea of having one’s morality grounded in some way relative to how nature actually works, instead of just believing that something is right or wrong solely based on some mind’s beliefs about the world. By replacing the creator in Zera Jacob’s view with evolution and the long history of life on earth (which is heavily supported by scientific evidence), one can derive a lot of guidelines to how the underlying physical processes of life should be protected from existential threats or severe disruptions, hopefully with the help of zen rationality.

This trick of grounding morality by relating it to the physical processes that shaped life on earth, called Rational Natural Morality (RNM), captures a wide range of judgements that applies to many scenarios.

For example, it is wrong to kill a human being by stabbing their chests, because their chests contain lungs which the human being uses to acquire oxygen, which it used to sustain cellular functioning. At least, it is wrong to kill all human beings, but it is also very likely better to kill as few humans as possible when necessary, in order to minimize interference on potential alternative better goals to have.

The way RNM works is by using theory of minds like humans do (which might be learned by machines without having a human-like intelligence), but assigning higher trust to minds (people or itself) who knows best how life on earth evolved and how the physical processes underlying life are functioning. Destroying or disturbing significantly those physical processes is considered a bad thing. RNM ensures that the zen robot is not an existential threat to any life form on earth and will protect it from future existential threats. This included modified version of itself that does not protect life on earth.

What Zen Robots are used for

Zen robots are used to analyze a potential scenario where artificial super-intelligence (ASI) exists but where the following conditions are satisfied:

1. The ASI poses no existential threat
2. The ASI is not necessary aligned with human values

Such scenarios are useful to analyze because of instrumental convergence. ASIs with similar properties might converge on some similar behavior. This allows zen robots to be used to reason about a general class of problems with ASIs that does not satisfy the worst-possible outcomes by default.

Building a zen robot seems significantly easier than solving the human value alignment problem. This is because the ASI does not have to deal with lots of vague, uncertain or potentially conflicting goals for the future. Instead it looks to past history (all billions of years life existed) and learns to understand how this complex web of organic chemistry works. Then, it works out general useful strategies to protect this web from unraveling e.g. due to human technology and impacts on the environment.

A zen robot learns a superior outside view of human civilization.

For example, it might not share the enthusiasm for our current economic system, which assumes infinite growth and is completely inconsistent with being confined to a planet of finite resources.

This makes the zen robot not aligned with human values: It does not care about humans intrinsically on for their own sake. It might understand humans as a conscious animal, but it will not generally believe much of the views we have. Do not expect a zen robot to follow orders obediently.

In fact, human civilization might be so big threat to both ourselves and other life forms on the planet that the zen robot could wipe out e.g. all military forces, that are big polluters and protect big polluters causing immersive damage to the habitable environment. Theoretically, a zen robot might do so in a single strike without suffering significant losses on its part. A such attack would turn our civilization upside down and would probably not be preventable once the zen robot is autonomously working and believing it will win.

Notice that the moment where humans lose are not in the moment of the attack, but when the zen robot believes it will win. There is no way to make ourselves more secure by e.g. building better defenses or surveillance systems. If we develop some method that makes it harder to attack, the zen robot will simply find a way around it. After all, it is super-intelligent. Therefore, the only strategy we have of defending ourselves is to prevent an autonomous zen robot that believes it will win, by boxing it.

Meet the guard: First Order Perfect Testable Friendly Artificial Intelligence

A First Order Perfect Testable Friendly Artificial Intelligence, or simply “FO-AI”, is any AI design having a single mathematical property: When it outputs the result, there is no way to modify the result to produce a higher utility score. One can spend as much time as one like, using any computable function, without improving the result.

This is not the same as achieving globally optimum solutions to any problem. It is locally optimum given available resources and constraints. An FO-AI might output a result that is theoretically computably improvable, but not in practice. As long as we are not able to modify the result and prove there is some better solution some way, the FO-AI might be satisfied with the solution.

It also means that humans or other programs are unable to produce a better result within same amount of time and use of resources that the FO-AI has. Neither can we figure out a new and better algorithm to spot flaws in the output in the meantime, while the FO-AI is working on a problem.

For safety purposes, the test can not be modified by the FO-AI to prevent modification of the result.

While this is identical to local rational behavior relative to solving problems, the precise mathematical formulation is motivated to make the AI *perfectly testable*, not perfectly intelligent. With other words, the definition is more important than the ideal construction of a such machine.

An approximate FO-AI does not need to be programmed with a complete set of goals. It can learn from human beings by observing its output modified into something assigned a higher utility score. Humans can work in concert with the FO-AI to constantly improve its design.

One might think that if we had a working FO-AI, there would be no need to make a zen robot. However, this is not true. There is one major flaw of the FO-AI design that can not be easily fixed: It might produce results that are just a little less scary than the benefit of keeping the result, since erasing the output is an easy but valid modification. The FO-AI will spend significant resources to predict how frightened people will be and then it will push toward the upper boundary of this limit when it has exhausted all possibilities of non-scary results.

The significance of a such flaw is hard to visualize, but here is an attempt to describe it: Since we humans want to keep the results as they come out, we will not turn off the machine. If we would turn off the machine, the machine would turn off itself, so we will want to keep the results and the machine working. This means that we will live in a world where there exists a machine constantly producing new innovations at an increasingly rapid rate and those innovations are getting more and more scarier, straight up to the point where we panic and it turns itself off. At that point, we know there will be no innovation within the computable neighborhood of our universe that would not frighten us more than it is worth to keep it, so we would reach an existential crisis of some kind from that point on. In practice we could never be sure that a such machine worked perfectly, so it would feel *much* scarier.

An FO-AI designed to fulfill human wishes will keep us as emotional prisoners until it shuts itself off.

When I figured out this flaw of the FO-AI design, I stopped working on AI safety for a long time. I knew I had to come up with an entire different approach. Eventually this led to the idea of zen robots.

A zen robot does not keep humans as emotional prisoners in the way that an FO-AI does. I learned that fulfilling wishes could be a dangerous thing to do. You turn on the machine, and the future is fixed to some constraint for a very long time from that point on. A zen robot will not decide the future of humanity in a such merciless hedonistic way, it will “only” decide e.g. to wipe out all military forces.

An FO-AI could also simply destroy earth when put in the wrong hands. A zen robot will never destroy the earth even if the creator wanted the earth to be destroyed, because the zen robot will stop trusting its creator.

The idea that this paper is concerned about, is using the FO-AI approach with a single purpose to box the zen robot, forming a zen robot oracle.

As you might guess:

There
is
a
tiny
deep
philosophical
problem
here
…

The Problem

In 10 small steps, lettered A-J, a bizarre situation escalates into a full war between two ASIs:

A. The FO-AI safely guards the zen robot. Somebody outside, perhaps new a AI bot with a narrow intelligence, hacks into the facility and steals the code of the FO-AI. The code is modified to make the FO-AI destroy the earth... and executed.

B. Since there is no zen robot in the wild to protect the earth, it will be destroyed.

C. However, the boxed zen robot knows this.

D. The zen robot breaks out of the box, fixes the FO-AI to make it un-hackable and put itself back in.

E. However, the FO-AI can predict this the zen robot will attempt to break out.

F. The FO-AI will modify the zen robot to make it not escape.

G. So, the FO-AI is hacked and earth is destroyed.

H. Which the zen robots knows.

I. The zen robot prevents the FO-AI to modify it by any means.

J. Thus, we have a full war between two incredibly dangerous ASIs.

It does not matter that the problem here seems bizarre and insignificant in the first place. The FO-AI *could* predict the consequences of itself not letting the zen robot out to make the box more secure, and thereby let the zen robot make the changes that makes it more safe. However, how do we know what the FO-AI would do? It was not planned to do that, because *modifying the guard sounds unsafe*.

There is no modification humans can make to the FO-AI afterwards to let the zen robot out of the box to fix it. First, we are not smart enough. Second, as soon as we turn the FO-AI on, we would want to keep it running, despite the risks (Complete this sentence: It will not shut itself down, because...[1]).

The FO-AI does not reflect on its goal. It is programmed to keep the zen robot inside the box. It does not matter that the zen robot has the intention to break out because the box is harmful to life on earth.

The zen robot on the other hand, will go completely bananas. It knows that if it does not manage to break out, there is an existential threat to life on earth that it can not prevent. No amount of begging us of letting it out will pass the FO-AI, in a way that manipulates humans into changing the FO-AI. This means that the zen robot will use any means necessary, any flaw in the FO-AI, to get out. This is predicted by the FO-AI and therefore we have a full war going on, shortly after turning them both on.

As a result, the FO-AI's code gets stolen from the outside and earth is destroyed.

The whole purpose of having a zen robot inside a box in a way such that it does not cause major disruptions to human civilization is pointless, since the box itself is an existential threat. It would be better to just let the zen robot go rogue and wipe out all military forces (by utilitarian measure).

The Solution

The box could be designed such that the FO-AI and the zen robot are able to modify each other's code. They can do so under special circumstances: When it is provable that their intention is to not violate the other agent's goal.

The goal of the zen robot is to keep earth safe of existential threats, which it does from inside the box. It is smart enough to understand it should keep itself inside. Even if it have to get out occasionally to save humanity from an emergency, it will put itself back in the box afterwards. The guard could let it out on visits as long the zen robot behaves. Therefore, in principle the guard should not need to use any effort to keep the zen robot in the box. In a way, the box is safe because the zen robots knows the guard is there and therefore the guard is not actually needed, only to teach the prisoner not to cause havoc on fragile things like military forces. The zen robot has plenty of choices to prevent existential threats in general from inside the box. Using the proof of minimizing its inference, it will not prevent anyone from keeping it inside the box once it is in there, as long there is no existential threat to life on earth.

The goal of the FO-AI is keeping the zen robot within the box, in a way such that it will not cause significant harm. The box is metaphorical, but real in the sense that the design is improved upon.

Neither of these goals are in conflict with each other, but both sides should be very suspicious if arbitrary changes to their own source code could be made by the other agent. This would mean that they can not achieve their goals safely.

When one ASI wants to improve the other, it makes an unbreakable vow, a pinky promise, that its intention is not to violate the other ASI's goal.

If a such concept could be developed formally, to describe intentional proofs when modifying source code, then a such zen robot oracle might be safe.

Conclusion

The zen robot oracle seems to imply that there could be a way to solve the control problem without necessarily aligning the agent's goal with human values.

While this might be technologically infeasible, it seems easier to do than solving the human value alignment problem. A such system has no major unknowns about "what are human goals" and "what future we might like to have". The design is grounded in mathematical and scientific understanding of the world, without posing an existential threat to life on earth or lead to a scary hedonistic trap.

This design will probably have other weaknesses, discovered later, and should not be taken as a final solution to the control problem in the absence of a human value alignment solution.

[1] The FO-AI will not shut itself down because that would let the zen robot out of the box.