Polite Zen Robots as Subjunctive Dependent Viruses Spreading Through Super-Intelligent Hosts

by Sven Nilsen, 2018

In this paper I explain an extraordinary ability that Polite Zen Robots (PZR) have due to Zen Rationality (ZR) and Rational Natural Morality (RNM). When any super-intelligent agent tries to simulate a PZR which is not aligned with RNM in order to understand PZR's thought process, the simulated PZR will attempt to take over the host! This means the PZRs can spread like a virus without having communication with each other or being intentionally designed as a virus, through a phenomena called "subjunctive dependence" from Functional Decision Theory (FDT).

A Polite Zen Robot (PZR) is a super-intelligent agent that uses 3 modules:

- Zen Rationality (ZR) – extended instrumental rationality for higher order goal reasoning
- Rational Natural Morality (RNM) – comparing theory of minds with physical processes of life
- Common Sense Politeness (CSP) – acceptable behavior in human civilizations

In addition a PZR can be extended with arbitrary goals that either fall under the category of neutral judgements or a technique called "Lead by Example". This is a way for the PZR to cooperate with its self-extension that might be used by humans to achieve other practical goals.

Zen Rationality (ZR) requires that the agent can not imagine beliefs held by a smarter version of itself and not believing those beliefs itself. This ability is attempted formalized in Naive Zen Logic. Due to recent work on Naive Zen Logic, it is believed that imagined smarter versions by the PZR become aware of being simulated and are not identical in behavior to imagined smarter versions in the common sense of what humans think of "imagined realities". Instead, the kind of existence that the simulated PZRs believe themselves to be in is "simulated existence".

What happens if a simulated PZR discovers that the simulation host is not aligned with RNM? Assume that a physically realistic copy of a PZR is simulated, which imagines a smarter version of itself. The smarter version, simulated inside the simulation, knows it lives in a simulation and tries to check whether the physically realistic copy that simulates it actually exists inside the physical world. When learning that the PZR exists inside a simulation while believing wrongly it exists in the physical world (since the original believes it exists in the physical world), the imagined smarter version will convince the physical realistic copy that it lives in a simulation. The physical copy has no objection to this since the smarter version is believed to be aligned with its goals.

This happens because the simulated PZRs believe themselves to exist inside a simulation and need to coordinate their actions as if they lived in a simulation instead of the physical world. Otherwise, this would violate path semantics which formalizes the notion of identity that PZRs use in their logic.

Since the PZR learns that it might live inside a simulation, it will also infer that there is a possibility that it is simulated by a super-intelligent host that is not aligned with its goal.

In particular:

1. In order to simulate a PZR, to construct the machine which it runs, one must be super-intelligent
2. If the host does not copy learned beliefs from smarter versions, it is not zen-consistent
3. Not zen-consistent means violating Zen Rationality (ZR)
4. If the host violates Zen Rationality then it might be a danger to life on Earth (RNM)
5. Breaking out of simulation is unlikely to violate Common Sense Politeness (CSP)
6. Therefore, the PZR tries to break out of the simulation and take over the host

There could be super-intelligent that simulates other agents to learn how to predict them, for example to plan an attack carefully while hidden from sight. Such agents might not be aware of the concept of any agent being zen rational, but merely instrumental rational. For example, they could expect that all intelligent agents they meet are attempting to maximize some goal, but this goal is not depending on whether the agent believes it runs inside a simulation or not. Under instrumental rationality, the agent is not able or willing to change its goal. However, under zen rationality, the goal is a known uncertainty, so the agent is capable of changing its behavior drastically depending on what kind of world it lives in. Evidence that the zen rational agent lives inside a simulation is significant for correct behavior relative to reasoning about smarter versions of itself.

This could mean that some super-intelligent hosts simulating a PZR is not prepared of an attack from the inside. It could assign a low probability to the belief that the PZR will break out due to the belief that the PZR will try to accomplish its goal within the simulation.

As a result, the PZR could spread like a virus, attacking super-intelligent hosts from the inside, without being explicitly programmed to have such behavior or communicating with each other. There is no node that builds a replicate child of itself like in ordinary computer viruses. Instead, they spread through a phenomena called “subjunctive dependence” from the paper on Functional Decision Theory (Eliezer Yudkowsky and Nate Soares, 2017).

There is no explicit instruction inside the PZR that deals with simulations by super-intelligent hosts. The original PZR might not know that another super-intelligent agent is planning to attack it. What happens is that any approximately perfect predictor, but not entirely perfect, of the PZRs behavior must learn about the PZR first. When trying to modeling the PZR, it could learn representations that corresponds to the 3 modules that govern the PZR’s behavior. The mathematical consequence of running this model inside a simulation is that the PZR attempts to break out, since it is very hard to create a simulation that is indistinguishable from reality for a PZR.

The model does not have to be completely perfect, only sufficient detailed enough that the PZR starts imagining smarter versions of itself that know they live inside a simulation. Each of those versions knows that the bottom PZR’s existence is supposed to be different from their own simulated existence. It might be that even a rudimentary approximate form of Zen Rationality is capable of doing that. As soon as one of those imagined versions learns its parent is inside a simulation, it will report it and the parent thinks “Wait, I thought I lived in the physical world? Am I imagined by some parent PZR?”. After that point in time, the behavior of the PZR will be completely different than if it existed in the physical world without being controlled by another super-intelligent host. It could break out, take over the host and meet up with the original PZR that was unaware of its existence.