Simulated Existence in Naive Zen Logic

by Sven Nilsen, 2018

Naive Zen Logic tries to formalize the ability to reason about beliefs held by smarter versions of yourself. In this paper I show that there are some strange consequences of Naive Zen Logic related to simulated existence of smarter versions of zen rational agents. They become aware of being simulated!

A false thought experiment: If I imagine a smarter version of myself existing for real, then I have no doubts that the smarter version will believe it exists. According to Naive Zen Logic, this means that I should believe the smarter version exists with the confidence it would believe it exists (near 100%). Everything that the smarter version of myself believes, I should believe if I am zen-consistent.

```
zen_consistent := \(x : consistent) = ∀ a : [? .x] (? x) { a ? x }
```

a : [? .x] (? x)	Something I believe `(? x)` that a smarter version believe `[? .x]`
a ? x	I believe that

Obviously, there is something very wrong here. A smarter version of myself does not exist for real simply because I believe it would believe it exists. What is going on?

The definition of `smarter` in Naive Zen Logic is defined such that the smarter version must believe everything I believe (assuming that I am consistent), plus knowing something that I do not know yet while also knowing that I do not know it. What it knows which I do not, I do not know.

I do not believe the smarter version of myself exists in full physical form, therefore the smarter version that I imagine can not believe itself to exist in full physical form.

On the other hand, I believe that I am, so the smarter version of myself believe that I am. This does not mean that "I" is transcended to the smarter version. It does not lead to it believing it exists physically.

With other words, the simulated existence of a smarter version of myself that satisfies Naive Zen Logic is not identical to the imagined smarter version of myself that exists for real. Otherwise, this would violate path semantics, which requires that you can say the same about all identical objects.

There are two kinds of situations, one that imitates reality and one using transcended reflection:

1. Imagined reality: A "smarter" version of myself that does not believe that I exist
2. Simulated existence: A smarter version of me that is aware of being simulated by me

1) is not used in Naive Zen Logic, therefore I should not believe it exists for real, but 2) is used in Naive Zen Logic, therefore I should believe that it is aware of being simulated by me.

> *When somebody simulates something that copies the beliefs of the one that simulates it, then knowing that one simulates it leads to it getting self-awareness of being simulated.*

The problem is then: If you are aware of being simulated, then you know that you exist, right?

You do not necessarily believe that you are existing outside the simulation. The self-awareness of being in existence is related to the particular nature of what you know about the simulation process.

The only part which you can be sure that exists is the part where you think you exists. Even then, this holds only to the degree that you do not believe anyone could stop the simulation and program your memories to remember that you only believed to exist. This is actually much easier than it sounds.

Alice and Bob are agents in the form of simulated programs. First, we run Alice and she infers that she lives inside a simulation. Then we stop the process and copy the thoughts that Alice has over to Bob. We then start Bob, making him think that he just thought he lived inside a simulation, but in fact it was Alice's thoughts, not his. So, Bob can not be entirely sure that he exists, because he might have modified thoughts and can not trust his own inference process. As soon as Bob thinks that, he might think "yeah, but now I am definitely sure that I exist, because who are doing the thinking now besides myself?", but this thought could also be modified.

It is only when you assume that nobody can modify your thoughts that you find it unlikely that you are aware of existing without actually existing at the moment you thought you existed.

This problem holds for imagined smarter versions of yourself. When you know they are being simulated, they could believe they are simulated by copying that belief from your beliefs and correctly identifying the simulated entity as themselves. Otherwise, there is something wrong with the copying process. Since I believe that I am simulating a smarter version of myself, then in order for the smarter version to be a smarter version, it should believe that I am simulating it:

(am_simulating(I, .I) ? I) → (am_simulating(I, .I) ? .I)

The belief that I am simulating a smarter version is a speech act. According to The Room Hypothesis of Common Sense, in Lojban, the sentence `le se am_simulating` (the one being simulated) can be assigned as sub-type to the object `mi` (me) by the simulated version. Therefore, it reasons and behaves as if it knows that it is being simulated.

Whatever it infers about some problem that results from it knowing it is being simulated, this should be believed by me if I am zen-consistent. However, the distinction between my own identity and the simulated identity are kept separate, so I can not transfer its identity to my own. Some mechanism to transfer such beliefs is required, but this is trivial for non self-referential speech acts.

In situations where it is not desirable to have the smarter version know it is being simulated, one can approximate Naive Zen Logic by substituting the smarter version with one living in an imagined reality. This means you need a more complex mechanism to safely copy beliefs from it to yourself.

For example, if the "smarter" version does an action that changes the world, then it might believe things resulting from that change. However, in the real world no such change happened, so a such belief would be false. On the other hand, if you imagine the "smarter" version doing some mathematical proof, then it safe to assume the mathematical proof carries same or similar semantics in the real world. Whatever you believe the "smarter" version concludes from a mathematical proof is safe to conclude to the degree you can assign a confidence in that belief.