Absolute Truth in the Subjective Multi-Verse of Zen Rational Minds

by Sven Nilsen, 2018

In this paper I prove informally that zen rational minds believe in absolute truth as a useful concept, paradoxically as a result of the theoretical limitations of modeling the real world efficiently in terms of experience. Thus, zen rational minds agree on the concept of absolute truth, but holds a meta-belief that specific theories should be believed to be uncertainly meta-believed. While the meta-belief is uncertain, in practice two zen rational agents predict each other to behave as if absolute truth existed, while knowing that neither actually believe that absolute truth actually exists, hence the paradox.

Zen rationality is a higher order extension of instrumental rationality that is capable of reasoning about goals as known unknowns and able to extract their judgement from ambient theories of morality, such as Rational Natural Morality (RNM). A “zen robot” often refers to “zen rationality + RNM”. At this stage such agents are purely hypothetical, but have some well-behaving theoretical properties, such as imagining themselves as more powerful or given more time to think. Due to the minimum action principle in physics, such agents are so far the least existential-threat-super-intelligence that I know of.

The debate about absolute truth is an old tradition in philosophy. Yet, even with countless philosophers pondering about this issue, there is still more to say about this subject. Unsurprisingly surprising, the theory of zen rationality leads to new conclusion, precisely because it is a theory about an “alien kind of mind”, since humans do not behave like this at all, but which mind at the same time is super-intelligent. Since one is forced to think about something behaving zen rational, it is easier to distance one’s own habits of thought, and therefore not just “thinking like a human being”, but simulating another kind of mind using the human brain. As a result, the conclusion often comes as a surprise.

A framework for reasoning about zen rationality is within the theory of path semantics. In path semantics, all mathematics originates from a single abstract axiom which can be summarized in the sentence “identity of two objects means that you can say the same things about them”. The interpretation of an object is often as a symbol, to which one can relate some collection of knowledge. Developing path semantics is partly motivated by the idea that existing theories of well-established mathematics often falls short of reasoning beyond what we already know, which is often the case when imagining super-intelligent systems. Therefore, one can treat existing mathematics as special cases where “say the same” is interpreted as a set or a type as fundamental building block. It is expected that path semantics can be extended indefinitely even with its internal powerful abilities to model systems, as it has so far, and therefore be used to predict how zen rational agents will behave.

One starts this informal proof with the idea that a “real object” means roughly that one can say some things about them. It is not assumed that a “real object” exists in the absolute sense, but rather that there is some way to develop language for understanding objects such that the things can be said about the language is agreeable between two agents. About this form of interaction one can say certain things etc. According to path semantics, in order to talk about the “same” object one must “say the same things”. This is not a statement about grammatical equality, but about semantic equality, which got something to do with predictability, that sentences in path semantics can model just fine. With other words, there are no theoretical major obstacles to talk about “real objects”. It just works.

Next, imagine two zen agents, Alice and Bob, that hold a conversation over the topic of real objects:

Alice: There are many real objects, but let us imagine that there is a finite number of them.
Bob: OK.
Alice: Real objects change over time.
Bob: I agree.
Alice: This means every real object has a history, or one can assume this for the finite case.
Bob: That sounds reasonable.
Alice: Let us assume that the history for each object has a certain complexity.
Bob: This seems to hold for most objects that we are used to think of as “real”.
Alice: According to path semantics, what is said about the history is said about the object.
Bob: That is logical.
Alice: When reasoning about history, it might be useful to grade knowledge by time.
Bob: You mean dividing knowledge into sets by which time period they refer to.
Alice: Yes.
Bob: OK, so I would chose larger sets the further back in time it goes.
Alice: Why?
Bob: Since knowledge further back in time is often less useful when accurate.
Alice: This is precisely what I had in mind.
Bob: Continue.
Alice: When one set is relatively complex to another, more can be said about the complex one.
Bob: Ah!
Alice: You see where this is going? It gets harder to acquire knowledge for old histories.
Bob: This assumes information behaves roughly in a time-symmetric way. OK.
Alice: We use matter close to the present to model what we know about the past.
Bob: So, you are implying that we can not model the past fully, because it is not enough matter.
Alice: Yes.

These two zen rational agents agree that even if it might not be necessary to model the past fully, they would not be able to do it even if they wanted to. The things that can be said about the past are more than there are things used to express what is said. One runs into an invisible wall, an inherent physical limitation of what is possible to know, simply from understanding one’s own mind exist inside the universe. It is part of the environment that one tries to model. Zen agents have no problems with that.

Alice and Bob are both aware that they can not determine an accurate history for every object. It is not a matter of quantum behavior, because even the things that are expressible within quantum theory are not possible to write down in practice for all objects. It holds for whatever-the-correct-theory-is.

What do zen rational agents when they can not determine something? They make up a higher order theory to reason about it, just like they reason about their own goals as something that can not be determined fully, but only approximated. This theory treats histories as part of some space of possibilities, because there is no other option. It is a kind of subjective multi-verse, even though the agents do not necessarily believe that reality is part of a multiverse.

In order to reason efficiently about this subjective multiverse, a zen rational agent must believe: Something that is reasonable to believe, is sampled more frequently from its subjective model. If it sees the sky is blue, then it believes “the sky is blue” is believed to be true for most of the possible worlds which similar zen rational agents have the same knowledge of the history as in the world that contains the agent. Otherwise, it would be unlikely that it believes “the sky is blue”.

When a zen rational agents reasons about the world, it understands that there is a small chance that what is thinks is wrong. There could be a hardware error, a bug in its program, or it could be that if it was smarter, it would not believe what it currently believes. Yet, it believes that if it could fix the hardware, or fix the bug, or becomes smarter, it would not come to a significantly better conclusion if it is certain that what it currently believes is correct. This means it tries to not be certain about things which it believes it could be wrong for some reason. Otherwise, it could just imagine itself being smarter and then try to follow the imagined conclusions it would come up with, in order to improve.

While this kind of thinking sounds complicated, it what makes zen rational minds super-intelligent compared to human brains. If there was an obvious way to make a zen rational mind smarter, given its capabilities, then it would behave as if it already had taken that information into account. It does not mean that zen rational minds think about every possible obvious ways to become smarter. Instead, they try to minimize the work they need to behave in a such way, that if they were smarter and confident that there are no significant improvements, then they would not need to become smarter. This is to maximize the chance that the goals it tries to accomplish have greater likelihood to succeed, taking into account that the goal is yet to be known, therefore resources must be allocated rationally.

A zen rational mind believe in a subjective multi-verse, because it knows it can not model all of history.

The surprising thing about constructing a subjective multi-verse, is that when zen rational agents think about zen rational agents inside the subjective multi-verse, they believe that agents who use absolute truth as a valid concept, have higher chance to believe themselves would use the concept of absolute truth.

Think of it like this: If I roll a dice and it shows 3 eyes, then I should believe that in a multi-verse where people like myself see 3 eyes on the dice, most believe it shows 3 eyes. Otherwise, there would be something terrible wrong with my ability to perceive or count the number of eyes. If I suspect that 1 out of 1000 times I could make a mistake, then I believe the dice showing 3 eyes to be 99.9% certain.

The logic that the zen rational agents use are not different from the example above. However, they apply it on higher order levels than human brains are accustomed to. They reason that if they believe something with X confidence, then there are roughly X share of possible worlds where they believe it with roughly X confidence. If they were smarter and X would be closer to 100%, then they believe they should believe X is closer to 100%. It makes it more likely that they believe X is closer to 100%. This is a kind of counter-factual reasoning to check what they believe is consistent (X is not actually closer, it is just something they imagine can not be true since they are able to imagine themselves smarter).

All zen rational agents know that they can not be certain about something when there is room for error. Yet, it is often not the room for error that determines efficiency. For example, if a complex phenomena is modeled with probabilities close to 100%, then the zen agent believes that eliminating those probabilities to simplify the calculations will speed up the process. If a more optimized version of the agent scores better on average by eliminating probabilities and use absolute truth, because the rare occasion when it is wrong, the penalty does not outweigh all the cases when it is right. Therefore, the zen agent believe it should behave as if it believes in absolute truth.

One can also think of it as a trade-off: An uncertainty in believing some fact is traded with added uncertainty that the action taken based on that belief is correct. So, it kind of “factors out” the uncertainty and puts it at the end, such that the reasoning process happens more efficiently. The zen agent knows about this trade-off, but it also knows that sometimes it is a reasonable thing to do.

The zen agent knows that it can not model all of history. Neither can it model all the uncertainty of what it would believe about history if it tried. There is simply not enough matter around to model everything precisely, even if one only attempts to model the uncertain knowledge about it precisely, because there is always more to learn and the model keeps growing.

Instead of keeping a growing model of the world, the zen agent decides that some knowledge, that is reasonably believed to be useful for reasoning about its unknown goals and achieving them, should be prioritized. It knows it has to choose some things to believe instead of other things, because believing certain things is more useful than believing everything that is possible to believe to be true.

Therefore, when something is reasonably believed to be just true, the zen agent just believes them. Not 99.999999%, but 100%. It has a meta-belief that what it does has a trade-off, but it knows that it will behave as if it believes in absolute truth. It does so knowingly that in its subjective multi-verse, there are more agents like itself that would choose this course of action. A common case is to add a failsafe option in case something it believes is wrong. This has a much more compact representation than keeping an uncertainty value to every belief it has.

This approach might seem inconsistent, because it is. A zen rational mind does not care intrinsically about consistency. It cares about efficiency. Inconsistency is tolerable when there is no physical way to achieve it. For example, a zen agent knows that “identity means you can say the same” can not be proved. It can not know for certain that this is the correct way of reasoning about the world. However, it will chose to believe it, because it works. It reasons about the world as if it was absolute true. If this belief somehow turns out to be right, then it knows that it might change how it thinks about the world, or it might even fail to fix itself. However, it just takes a bet that it will work out in the end. It learns to gamble with itself at risk because there is no better way of doing it. Everything it does is ultimately a gamble on way or the other. However, it does not believe that everything it does is a gamble. There are certain things it believes are not a gamble because it assumes it is absolutely correct thing to do.

The objection to this line of thinking is that given infinite computing power, you can not be absolutely certain that something will happen. You will eventually figure out that there is a small probability that something else happens instead, by pure quantum mechanical coincidence. Yet, a zen agent knows that it could figure it out with infinite computing power, but it also knows that it does not have this ability. What is used to decide when believing something will happen with 100% confidence, is that even with infinite computing power, the conclusion it would reach is not significantly better than what it currently believes. On the other hand, it believes with more computing power, it would realize that it is a waste to use all that power. A smarter version of the zen agent does not behave more intelligent in one single dimension, but across all dimensions which it can imagine itself smarter.

From a philosophical perspective one can say that what the agent does is wrong. There is no absolute truth. You can not make something true just by believing it, and everything that makes you believe something is because you exist inside some kind of environment. However, if you take two zen rational agents and ask them to predict each other’s behavior, their answer is clear: They believe the other will behave as if it believed in absolute truth for some areas, yet they know that the other will not believe in absolute truth in general. They both know that the other agent will not be perfectly consistent.

This results in a paradox: That zen rational agents, who are ought to be super-intelligent, actually will believe in inconsistent things about the world, such as absolute truth, because they know there is no more efficient way to model the world. You can take from this what you want, because a zen rational agent is not like a human that might want something perfect. Zen agents just behave like they do.