Rationality: From Al to Zombies
Eliezer Yudkowsky

Rationality from AI to Zombies – Semantic Stopsigns

This is all that is left – just a fragment from the book, to give you a taste.

And the child asked:

Q: Where did this rock come from?

A: I chipped it off the big boulder, at the center of the village.

Q: Where did the boulder come from?

A: It probably rolled off the huge mountain that towers over our village.

Q: Where did the mountain come from?

A: The same place as all stone: it is the bones of Ymir, the primordial giant.

Q: Where did the primordial giant, Ymir, come from?

A: From the great abyss, Ginnungagap.

Q: Where did the great abyss, Ginnungagap, come from?

A: Never ask that question.

Consider the seeming paradox of the First Cause. Science has traced events back to the Big Bang, but why did the Big Bang happen? It's all well and good to say that the zero of time begins at the Big Bang— that there is nothing before the Big Bang in the ordinary flow of minutes and hours. But saying this presumes our physical law, which itself appears highly structured; it calls out for explanation. Where did the physical laws come from? You could say that we're all a computer simulation, but then the computer simulation is running on some other world's laws of physics— where did those laws of physics come from?

At this point, some people say, "God!"

What could possibly make anyone, even a highly religious person, think this even helped answer the paradox of the First Cause? Why wouldn't you automatically ask, "Where did God come from?" Saying "God is uncaused" or "God created Himself" leaves us in exactly the same position as "Time began with the Big Bang." We just ask why the whole metasystem exists in the first place, or why some events but not others are allowed to be uncaused.

My purpose here is not to discuss the seeming paradox of the First Cause, but to ask why anyone would think "God!" could resolve the paradox. Saying "God!" is a way of belonging to a tribe, which gives people a motive to say it as often as possible— some people even say it for questions like "Why did this hurricane strike New Orleans?" Even so, you'd hope people would notice that on the particular puzzle of the First Cause, saying "God!" doesn't help. It doesn't make the paradox seem any less paradoxical even if true. How could anyone not notice this?

Jonathan Wallace suggested that “God!” functions as a semantic stopsign— that it isn’t a propositional assertion, so much as a cognitive traffic signal: do not think past this point. Saying “God!” doesn’t so much resolve the paradox, as put up a cognitive traffic signal to halt the obvious continuation of the question-and-answer chain.

Of course you’d never do that, being a good and proper atheist, right? But “God!” isn’t the only semantic stopsign, just the obvious first example.

The transhuman technologies— molecular nanotechnology, advanced biotech, genetech, Artificial Intelligence, et cetera— pose tough policy questions. What kind of role, if any, should a government take in supervising a parent’s choice of genes for their child? Could parents deliberately choose genes for schizophrenia? If enhancing a child’s intelligence is expensive, should governments help ensure access, to prevent the emergence of a cognitive elite? You can propose various institutions to answer these policy questions— for example, that private charities should provide financial aid for intelligence enhancement— but the obvious next question is, “Will this institution be effective?” If we rely on product liability lawsuits to prevent corporations from building harmful nanotech, will that really work?

I know someone whose answer to every one of these questions is “Liberal democracy!” That’s it. That’s his answer. If you ask the obvious question of “How well have liberal democracies performed, historically, on problems this tricky?” or “What if liberal democracy does something stupid?” then you’re an autocrat, or libertopian, or otherwise a very very bad person. No one is allowed to question democracy.

I once called this kind of thinking “the divine right of democracy.” But it is more precise to say that “Democracy!” functioned for him as a semantic stopsign. If anyone had said to him “Turn it over to the Coca-Cola corporation!,” he would have asked the obvious next questions: “Why? What will the Coca-Cola corporation do about it? Why should we trust them? Have they done well in the past on equally tricky problems?”

Or suppose that someone says “Mexican-Americans are plotting to remove all the oxygen in Earth’s atmosphere.” You’d probably ask, “Why would they do that? Don’t Mexican-Americans have to breathe too? Do Mexican-Americans even function as a unified conspiracy?” If you don’t ask these obvious next questions when someone says, “Corporations are plotting to remove Earth’s oxygen,” then “Corporations!” functions for you as a semantic stopsign.

Be careful here not to create a new generic counterargument against things you don’t like—“ Oh, it’s just a stopsign!” No word is a stopsign of itself; the question is whether a word has that effect on a particular person. Having strong emotions about something doesn’t qualify it as a stopsign. I’m not exactly fond of terrorists or fearful of private property; that doesn’t mean “Terrorists!” or “Capitalism!” are cognitive traffic signals unto me. (The word “intelligence” did once have that effect on me, though no longer.) What distinguishes a semantic stopsign is failure to consider the obvious next question.

Yudkowsky, Eliezer (2015-03-11). Rationality: From AI to Zombies (Kindle Locations 2355-2405). Machine Intelligence Research Institute. Kindle Edition.