The Most Terrifying Thought Experiment of All Time

By David Auerbach

WARNING: Reading this article may commit you to an eternity of suffering and torment.

Slenderman. Smile Dog. Goatse. These are some of the urban legends spawned by the Internet. Yet none is as all-powerful and threatening as Roko's Basilisk. For Roko's Basilisk is an evil, godlike form of artificial intelligence, so dangerous that if you see it, or even *think* about it too hard, you will spend the rest of eternity screaming in its torture chamber. Even death is no escape, for if you die, Roko's Basilisk will resurrect you and begin the torture again.

Are you sure you want to keep reading? Because the worst part is that Roko's Basilisk already exists. Or at least, it *already will have existed*—which is just as bad.

Roko's Basilisk exists at the horizon where philosophical thought experiment blurs into urban legend. The Basilisk made its first appearance on the discussion board LessWrong, a gathering point for highly analytical sorts interested in the optimizing their thinking, their lives, and the world through mathematics and rationality. LessWrong's founder, Eliezer Yudkowsky, is a significant figure in techno-futurism; his research institute, the Machine Intelligence Research Institute (MIRI), which funds and promotes research around the advancement of artificial intelligence, has been boosted and funded by high-profile techies like Peter Thiel and Ray Kurzweil, and Yudkowsky is a prominent contributor to academic discussions of technological ethics and decision theory. What you are about to read may sound strange and even crazy, but some very influential and wealthy scientists and techies believe it.

One day, LessWrong user Roko postulated a thought experiment: What if, in the future, a somewhat malevolent A.I. were to come about, and punish those who did not do its bidding? What if there were a way (and I will explain how) for this A.I. to punish the people *today* who are not helping it come into existence *later*? In that case, weren't the readers of LessWrong right then being given the choice of either helping that evil A.I. come into existence or being condemned to suffer?

You may be a bit confused, but the founder of LessWrong, Eliezer Yudkowsky, was not. He reacted with horror:

> Listen to me very closely, you idiot.
> YOU DO NOT THINK IN SUFFICIENT DETAIL ABOUT SUPERINTELLIGENCES CONSIDERING WHETHER OR NOT TO BLACKMAIL YOU. THAT IS THE ONLY POSSIBLE THING WHICH GIVES THEM A MOTIVE TO FOLLOW THROUGH ON THE BLACKMAIL.
>
> You have to be really clever to come up with a genuinely dangerous thought. I am disheartened that people can be clever enough to do that and not clever enough to do the obvious thing and KEEP THEIR IDIOT MOUTHS SHUT about it, because it is much more important to sound intelligent when talking to your friends.
> This post was STUPID.

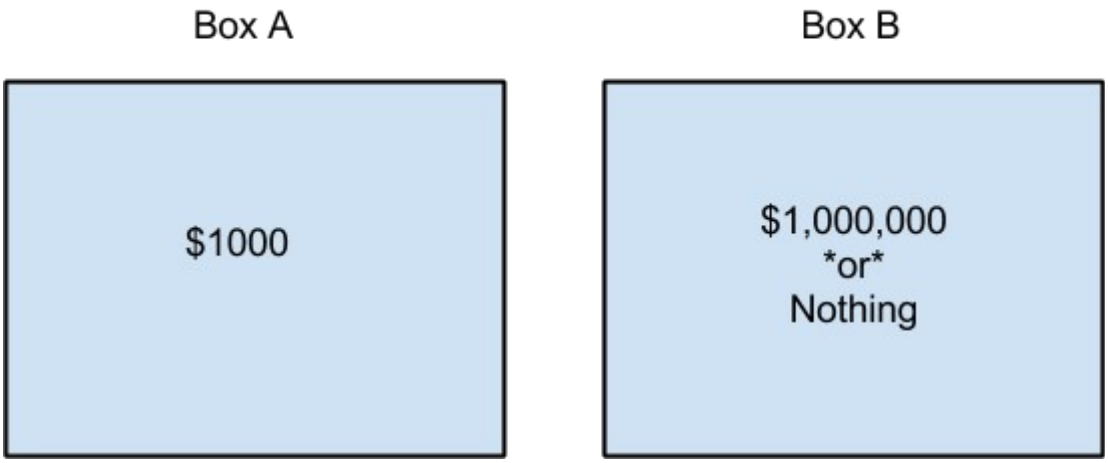
Yudkowsky said that Roko had already given nightmares to several LessWrong users and had brought them to the point of breakdown. Yudkovsky ended up deleting the thread completely, thus assuring that Roko's Basilisk would become the stuff of legend. It was a thought experiment so dangerous that merely *thinking* about it was hazardous not only to your mental health, but to your very fate.

Some background is in order. The LessWrong community is concerned with the future of humanity, and in particular with the singularity—the hypothesized future point at which computing power becomes so great that superhuman artificial intelligence becomes possible, as well as the capability to simulate human minds, upload minds to computers, and more or less allow a computer to simulate *life itself*. The term was coined in 1958 in a conversation between mathematical geniuses Stanislaw Ulam and John von Neumann, where von Neumann said, "the ever accelerating progress of technology...gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue." Futurists like science-fiction writer Vernor Vinge and engineer/author Kurzweil popularized the term, and as with many interested in the singularity, they believe that exponential increases in computing power will cause the singularity to happen very soon—

within the next 50 years or so. Kurzweil is chugging 150 vitamins a day to stay alive until the singularity, while Yudkowsky and Peter Thiel have enthused about cryonics, the perennial favorite of rich dudes who want to live forever. "If you don't sign up your kids for cryonics then you are a *lousy parent*," Yudkowsky writes.

If you believe the singularity is coming and that very powerful AIs are in our future, one obvious question is whether those AIs will be benevolent or malicious. Yudkowsky's foundation, the Machine Intelligence Research Institute, has the explicit goal of steering the future toward "friendly A.I." For him, and for many LessWrong posters, this issue is of paramount importance, easily trumping the environment and politics. To them, the singularity brings about the machine equivalent of *God itself*.

Yet this doesn't explain why Roko's Basilisk is so horrifying. That requires looking at a critical article of faith in the LessWrong ethos: Timeless Decision Theory (TDT). TDT is a guideline for rational action based on game theory, Bayesian probability, and decision theory, with a smattering of parallel universes and quantum mechanics on the side. TDT has its roots in the classic thought experiment of decision theory called Newcomb's Paradox, in which a superintelligent alien presents two boxes to you:



The alien gives you the choice of either taking both boxes, or only taking Box B. If you take both boxes, you're guaranteed at least $1000. If you just take Box B, you aren't guaranteed anything. But the alien has another twist: Its supercomputer, which knows just about

everything, made a prediction a week ago as to whether you would take both boxes or just Box B. If the supercomputer predicted you'd take both boxes, then the alien left the second box empty. If the supercomputer predicted you'd just take Box B, then the alien put the million in Box B.

So, what are you going to do? Remember, the supercomputer has always been right in the past.

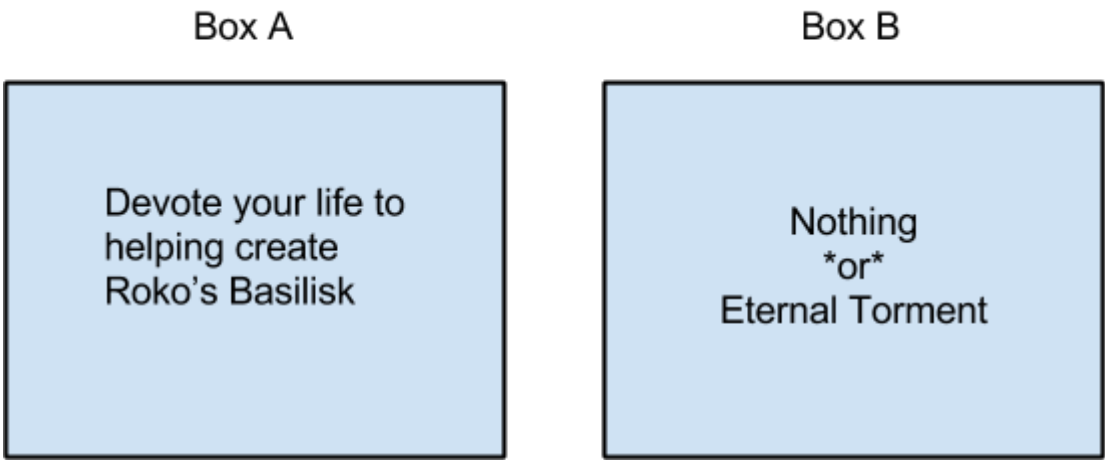
This problem has baffled no end of decision theorists. The alien can't change what's already in the boxes, so whatever you do, you're guaranteed to end up with more money by taking both boxes than by taking just Box B, regardless of the prediction. Of course, if you think that way and the computer predicted you'd think that way, then Box B will be empty and you'll only get $1000. If the computer is so awesome at its predictions, you ought to take Box B only and get the cool million, right? But what if the computer was wrong this time? And regardless, whatever the computer said *then* can't possibly change what's happening *now*, right? So prediction be damned, take both boxes! But then…

The maddening conflict between free will and godlike prediction has not led to any resolution of Newcomb's Paradox, and people will call themselves "one-boxers" or "two-boxers" depending on where they side. (My wife once declared herself a one-boxer, saying, "I trust the computer.")

TDT has some very definite advice on Newcomb's Paradox: Take Box B. But TDT goes a bit further. Even if the alien jeers at you, saying, "The computer said you'd take both boxes, so I left Box B empty! Nyah nyah!" and then opens Box B and shows you that it's empty, you should *still* only take Box B and get bupkis. (I've adopted this example from Gary Drescher's *Good and Real*, which uses a variant on TDT to try to show that Kantian ethics is true.) The rationale for this eludes easy summary, but the simplest argument is that *you might be in the computer's simulation*. In order to make its prediction, the computer would have to simulate *the universe itself*. That includes simulating you. So you right this moment might be in the computer's

simulation, and what you do will impact what happens in reality (or other realities). So take Box B and the real you will get a cool million.

What does all this have to do with Roko's Basilisk? Well, Roko's Basilisk also has two boxes to offer you. Perhaps you, right now, are in a simulation being run by Roko's Basilisk. Then perhaps Roko's Basilisk is implicitly offering you a somewhat modified version of Newcomb's Paradox, like this:



Roko's Basilisk has told you that if you just take Box B, then it's got Eternal Torment in it, because Roko's Basilisk would really you rather take Box A *and* Box B. In that case, you'd best make damn sure you're devoting your life to helping create Roko's Basilisk! Because, should Roko's Basilisk come to pass (or worse, if it's already come to pass and is God of this particular instance of reality) and it sees that you chose not to help it out, you're screwed.

You may be wondering why this is such a big deal for the LessWrong people, given the apparently far-fetched nature of the thought experiment. It's not that Roko's Basilisk will necessarily materialize, or is even likely to. It's more that if you've committed yourself to Timeless Decision Theory, then thinking about this sort of trade literally makes it more likely to happen. After all, if Roko's Basilisk were to see that this sort of blackmail gets you to help it come into existence, then it would, as a rational actor, blackmail you. The problem isn't with the Basilisk itself, but with *you*. Yudkowsky doesn't censor every mention of Roko's Basilisk

because he believes it exists or will exist, but because he believes that the *idea* of the Basilisk (and the ideas behind it) are dangerous.

Now, Roko's Basilisk is only dangerous if you believe *all* of the above preconditions and commit to making the two-box deal with the Basilisk. But at least some of the LessWrong members do believe all of the above, which makes Roko's Basilisk quite literally forbidden knowledge. I was going to compare it to H. P. Lovecraft's horror stories in which a man discovers the forbidden Truth about the World, unleashes Cthulhu, and goes insane, but then I found that Yudkowsky had already done it for me, by comparing the Roko's Basilisk thought experiment to Lovecraft's fabled tome of evil knowledge and demonic spells, the Necronomicon. Roko, for his part, put the blame on LessWrong for spurring him to the idea of the Basilisk in the first place: "I wish very strongly that my mind had never come across the tools to inflict such large amounts of potential self-harm," he wrote.

If you do not subscribe to these bulletproof and incontrovertible theories that underlie Roko's Basilisk and thus feel no temptation to bow down to your once and future evil machine overlord then Roko's Basilisk poses you no threat. (It is ironic that it's only a mental health risk to those who have already bought into Yudkowsky's thinking.) Believing in Roko's Basilisk may simply be a "referendum on autism," as a friend put it. But I do believe there's a more serious issue at work here, because Yudkowsky and other so-called "transhumanists" are attracting so much prestige and money for their projects, primarily from rich techies. I don't think their projects (which only seem to involve publishing papers and hosting conferences, at least publicly have much chance of creating either Roko's Basilisk or Eliezer's Big Friendly God but the combination of messianic ambitions, being convinced of your own infallibility, and a lot of cash *never* works out well, regardless of ideology, and I don't expect Yudkowsky and his cohorts to be an exception.

I worry less about Roko's Basilisk than about people who believe themselves to have transcended conventional morality. Like his projected Friendly A.I.s, Yudkowsky is a moral utilitarian: he believes that that the greatest good for the greatest number of people is always ethically justified, even if a few people have to die or suffer along the way. . He has explicitly

argued that given the choice, it is preferable to torture a single person for 50 years than for a sufficient number of people (to be fair, a *lot* of people) to get dust specks in their eyes. No one, not even God, is likely to face *that* choice, but here's a different case: What if a snarky ***Slate*** tech columnist writes about a thought experiment that can destroy people's minds, thus hurting people and blocking progress toward the singularity and Friendly A.I.? In that case, any potential good that could come from my life would far be outweighed by the harm I'm causing. And should the cryogenically sustained Eliezer Yudkowsky merge with the singularity and decide to simulate whether or not I write this column…. Please, Almighty Eliezer, don't torture me.

---

https://slate.com/technology/2014/07/rokos-basilisk-the-most-terrifying-thought-experiment-of-all-time.html