AI Defense in Depth: A Layman's Guide

What is the problem?

First post in sequence AI Defense in Depth: A Layman's Guide

AI Risk for Cavemen

A parable:

> A caveman is looking at a group of cavemen shamans performing a strange ritual that summons many strange, inscrutable, and increasingly useful daemons. Some of the shamans say they're going to eventually summon a daemon so powerful that it could solve all problems, now and forever. But they warn that if this *same daemon* has no heart, just like every other daemon that has ever been summoned, it might *devour the whole world* because of unfathomable daemon urges. Other shamans say that this cannot happen, that the ritual cannot bring a daemon so powerful into the world. Besides, some of these second group of shamans claim, even a daemon so powerful would not eat the world because *it would necessarily care about cavemen, in spite of being a daemon, due to being so powerful.*
>
> The first group of shamans points out that power does not require a heart, and that nobody has any idea how to give a daemon a heart. The second group of shamans ignores them. All the shamans continue the ritual. Increasingly more powerful daemons are being summoned, and the concerned group warns the World-Devourer could show up much sooner than anyone had thought.
>
> The caveman wonders which of the shamans are right. He also wonders why the ritual continues as is. It’s like no one listens to the concerned shamans. If they’re right the world will be eaten! What are they thinking? The caveman wonders whether the shamans can be trusted.

A Primer on Long-term Thinking

> Any sufficiently advanced technology is indistinguishable from magic.
> -- Arthur C. Clarke

Back here in our apparently wholly natural world, we have invented something *like* magic. It is called *science*, and on the whole it has worked out pretty well so far.

Climate change is a very thorny problem, and it is not looking like we will manage to fix it. While it will hurt, a lot, it is unlikely to kill us. The best shamanic estimate is 1/1000 chance of human extinction per century.

But even so. Not great.

What if we had taken the problem seriously when we first acknowledged it, we wonder? Could we have prevented this? Maybe. But even then, it would have been painful. In the 60s, greenhouse gas (GHG) emitting technologies were already thoroughly embedded into every aspect of civilization. It would have been better to have handled it then, but it would still have been a monumental task.

What if we had taken it seriously in *the 19th century*? Some shamans at the time were already theorizing about it. Could it have been stopped then? That would definitely have been the easiest time in which to stop it, before GHG emitting tech had become too widespread, but when there was a plausible route from the then current technologies to planetary disaster.

It would still have been a massive political struggle. GHG tech is too *powerful*, too *useful*, to easily prevent everyone from using it.

But it’s not the 19th century. Here in the 21st century, we do pull off regulating scientific endeavors, as the lack of genetically engineered humans anywhere on the planet shows.

It *is* the 19th century in another important sense though: there is another planetary disaster looming, and we will never be better positioned to tackle it than now.

AI Risk for Laymen

The looming planetary disaster is unaligned AI. What does that mean? It just means the proliferation of AI systems that do not care at all for human concerns. Like every other piece of software ever written.

But unlike any other piece of software, AI technologies have the potential of outmatching humans across every relevant domain, such that humanity would no longer have control over their future, outcompeted by software, at the mercy of software, just like gorillas and lions are at our mercy.

Like with gorillas and lions, AI wouldn't need to harbor any "ill-will" to humanity to harm or eradicate us: we don't hate gorillas, we just don't care enough whether our actions do or do not harm them. Since intellect is *power*, and we're so much smarter than gorillas, we're also much more *powerful* than them, such that the gorillas can ultimately do nothing when our actions harm them.

"But wait!" you might object. "It's impossible for AI to do that to us! We know exactly what intelligence is and what it can or cannot do!"

I say "Huh?".

You blink.

Exactly.

"But hang on. We don't have any reason to think we could develop an AI so powerful."

Well, you may think that, but you and me are laymen. Here is what leading AI expert Stuart Russell has to say on the matter:

> The risks of superintelligence can also be dismissed by arguing that superintelligence cannot be achieved. These claims are not new, but it is surprising now to see AI researchers themselves claiming that such AI is impossible. For example, a major report from the AI100 organization, "Artificial Intelligence and Life in 2030 [PDF]", includes the following claim: "Unlike in the movies, there is no race of superhuman robots on the horizon or probably even possible."
>
> To my knowledge, this is the first time that serious AI researchers have publicly espoused the view that human-level or superhuman AI is impossible—and this in the middle of a period of extremely rapid progress in AI research, when barrier after barrier is being breached. It's as if a group of leading cancer biologists announced that they had been fooling us all along: They've always known that there will never be a cure for cancer.
>
> What could have motivated such a volte-face? The report provides no arguments or evidence whatever. (**Indeed, what evidence could there be that no physically possible arrangement of atoms outperforms the human brain?**) I suspect that the main reason is tribalism—the instinct to circle the wagons against what are perceived to be "attacks" on AI. It seems odd, however, to perceive the claim that superintelligent AI is possible as an attack on AI, and even odder to defend AI by saying that AI will never succeed in its goals. We cannot insure against future **catastrophe** simply by betting against human ingenuity.

"But Stuart Russell is just one guy (who's a computer science professor at Berkeley, neurosurgery professor at UCSF, DARPA advisor, and author of the leading textbook on AI)! Facebook's Yann LeCun disagrees!"

[Shane Legg, co-founder of Google's DeepMind, is with Russell](#)

[Andrew Ng (roles have included Chief Scientist at Baidu, Google Brain lead, Coursera founder) is with me!](#)

We can keep playing Pokémon with scientists for quite a while (go [Steve Omohundro](#)!).

But the very fact that we can is the problem. If we want to know what the collective AI community, academia, industry, and government, thinks of AI safety, *there is no consensus to gesture to!*

And ordinarily, this would not matter. Why should mere cavemen care about the arcane disputes of shamans? They'll figure it out.

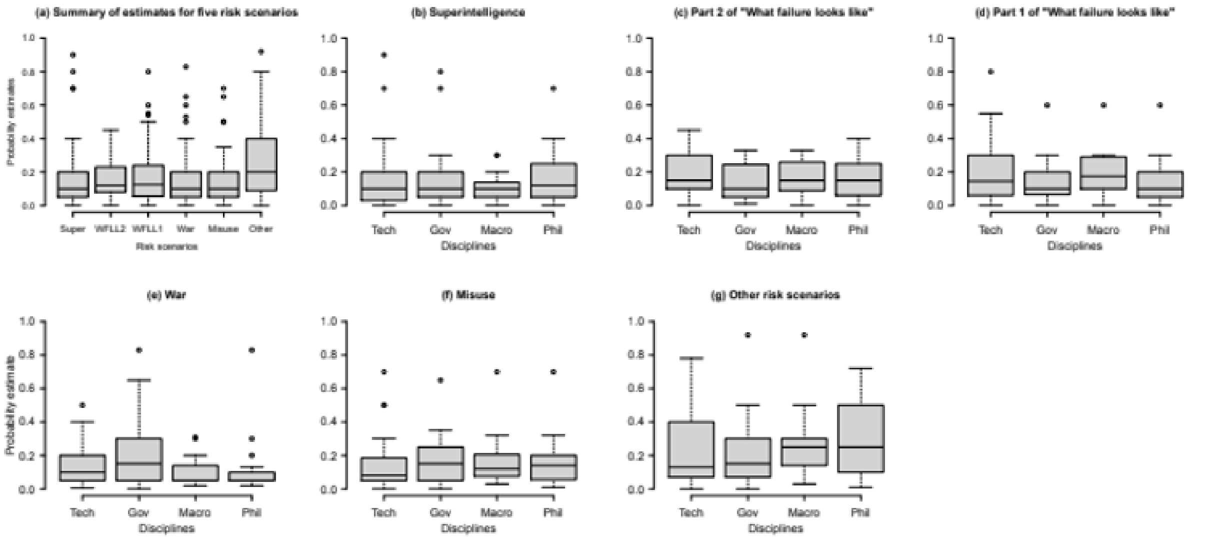
Well. Remember the estimate above of 1/1000 odds of human extinction from climate change? *It's 1/10 odds of extinction from unaligned AI*.

"Bull. No way that Toby Ord guy is right."

How would you know? How do you know you're not standing at the door of the control room at Chernobyl the night of the disaster, hearing the technicians bickering while [Anatoly Dyatlov](#) downplays the risks, downplays *even the idea that there are any risks at all*?

But forget Ord. Here is what the community thinks of the risk of AI:

Figure 2
Conditional probability estimates for five scenarios in which AI causes an existential catastrophe



Note. **(a)** summarises the conditional probability estimates for five scenarios in which AI causes an existential catastrophe, conditional on an existential catastrophe due to AI having occurred (*Super* = "Superintelligence"; *WFLL2* = Part 2 of "What failure looks like"; *WFLL1* = Part 1 of "What failure looks like"; *War* = some kind of war exacerbated by AI; *Misuse* = intentional misuse of AI; for more information on the scenarios, see Table 2). **(b)–(g)** give a breakdown of responses by discipline (*Tech* = technical AI safety; *Gov* = AI governance; *Macro* = macrostrategy and cause prioritization; *Phil* = philosophy). Center lines show the medians; the limits of the grey boxes indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles; outliers are represented by dots.

I can kinda make some sense out of that, but let's defer to a [shaman](#):

1. Even people working in the field of aligning AIs mostly assign "low" probability (~10%) that unaligned AI will result in human extinction

2. While some people are still concerned about the superintelligence scenario, concerns have diversified a lot over the past few years

3. People working in the field don't have a specific unified picture of what *will* go wrong

That's all very calm. Very sedate. But let's bring it back to Chernobyl at the night of the fatal test. You are a guard outside the control room, hearing some bickering. You show some initiative and boldly enter the room. "Everything okay?" you ask.

One of the technicians replies:

> We think there are 10% odds that if we continue the test the reactor will explode.

Another says:

> Some of us here are worried about an explosion, but others think some other bad stuff is more likely to happen.

And another:

> Bottom line, we are just not sure what will go wrong, but we think something will.

Dyatlov says:

> Worrying about an RBMK reactor explosion is like worrying about the Sun turning into a red giant and swallowing the Earth. Continue the test.

The guard, being a sensible person who defers to domain experts when it comes to their subject matter does... what? What does the guard do? *What do you do when experts bicker over what appears to be a life or death matter?*

I don't know. But for starters, it sounds like some of the experts don't belong in the room. You should stay in the room because technical experts don’t handle high pressure, low-information situations too well, being used to having the time and leisure to dot every i and cross every t before reaching a conclusion. We have evidence of that. If they could handle them, Chernobyl would not have happened, *because the technicians would just have beaten Dyatlov up and tossed him out of the room*. If they could handle them, Roger Boisjoly would have found a way to prevent the Challenger launch, *even if he had to scream or bloody some manager noses!*

This guide starts from there. We start by determining who are the Dyatlovs in the room and evicting them, even if it takes **extraordinary, heroic action**. Because what some of the experts are saying is that an AI Chernobyl may not be survivable.

In the meantime, tell other laymen!

Substack

Twitter

Could you have stopped Chernobyl?

...or would you have needed a PhD for that?

It would appear the inaugural post caused some (off-LW) consternation! It would, after all, be a tragedy if the guard in our Chernobyl thought experiment overreacted and just unloaded his Kalashnikov on everyone in the room and the control panels as well.

And yet, we must contend with the issue that if the guard had simply deposed the leading expert in the room, perhaps the Chernobyl disaster would have been averted.

So the question must be asked: can laymen do anything about expert failures? We shall look at some man-made disasters, starting of course, with Chernobyl itself.

Chernobyl

One way for problems to surface

To restate the thought experiment: the night of the Chernobyl disaster, you are a guard standing outside the control room. You hear increasingly heated bickering and decide to enter and see what's going on, perhaps right as Dyatlov proclaims there is no rule. You, as the guard, would immediately be placed in the position of having to choose to either listen to the technicians, at least the ones who speak up and tell you something is wrong with the reactor and the test must be stopped, or Dyatlov, who tells you nothing is wrong and the test must continue, and to toss the recalcitrant technicians into the infirmary.

If you listen to Dyatlov, the Chernobyl disaster unfolds just the same as it did in history.

If you listen to the technicians and wind up tossing *Dyatlov* in the infirmary, what happens? Well, perhaps the technicians manage to fix the reactor. Perhaps they don't. But if they do, they won't get a medal. Powerful interests were invested in that test being completed on that night, and some unintelligible techno-gibberish from the technicians will not necessarily convince them that a disaster was narrowly averted. Heads will roll, and not the guilty ones.

This has broader implications that will be addressed later on, but while tossing Dyatlov in the infirmary would not have been enough to really prevent disaster, it seems like it would have worked on that night. To argue that the solution is not actually as simple as evicting Dyatlov is not the same as saying that Dyatlov should not have been evicted: to think something is seriously wrong and yet obey is hopelessly akratic.

But for now we move to a scenario more salvageable by individuals.

The Challenger

Roger Boisjoly, Challenger warner

The Challenger disaster, like Chernobyl, was not unforeseen. Morton-Thiokol engineer Roger Boisjoly, had raised red flags with the faulty O-rings that led to the loss of the shuttle and the deaths of seven people as early as six months before the disaster. For most of those six months, that warning, as well as those of other engineers went unheeded. Eventually, a task force was convened to find a solution, but it quickly became apparent the task force was a toothless, do-nothing committee.

The situation was such that Eliezer Yudkowsky, leading figure in AI safety, held up the Challenger as a failure that showcases hindsight bias, the mistaken belief that a past event was more predictable than it actually was:

> Viewing history through the lens of hindsight, we vastly underestimate the cost of preventing catastrophe. In 1986, the space shuttle Challenger exploded for reasons eventually traced to an O-ring losing flexibility at low temperature (Rogers et al. 1986). There were warning signs of a problem with the O-rings. But preventing the Challenger disaster would have required, not attending to the problem with the O-rings, but attending to every warning sign which seemed as severe as the O-ring problem, without benefit of hindsight.

This is wrong. There were no other warning signs as severe as the O-rings. Nothing else resulted in an engineer growing this heated the day before launch (from the obituary already linked above):

> But it was one night and one moment that stood out. On the night of Jan. 27, 1986, Mr. Boisjoly and four other Thiokol engineers used a teleconference with NASA to press the case for delaying the next day’s launching because of the cold. At one point, Mr. Boisjoly said, he slapped down photos showing the damage cold temperatures had caused to an earlier shuttle. It had lifted off on a cold day, but not this cold.
>
> “How the hell can you ignore this?” he demanded.

How the hell indeed. In an unprecedented turn, in that meeting NASA management was blithe enough to reject an explicit no-go recommendation from Morton-Thiokol management:

> During the go/no-go telephone conference with NASA management the night before the launch, Morton Thiokol notified NASA of their recommendation to postpone. NASA officials strongly questioned the recommendations, and asked (some say pressured) Morton Thiokol to reverse their decision.
>
> The Morton Thiokol managers asked for a few minutes off the phone to discuss their final position again. The management team held a meeting from which the engineering team, including Boisjoly and others, were deliberately excluded. The Morton Thiokol managers advised NASA that their data was inconclusive. NASA asked if there were objections. Hearing none, NASA decided to launch the STS-51-L Challenger mission.
>
> Historians have noted that this was the first time NASA had ever launched a mission after having received an explicit no-go recommendation from a major contractor, and that questioning the recommendation and asking for a reconsideration was highly unusual. Many have also noted that the sharp questioning of the no-go recommendation stands out in contrast to the immediate and unquestioning acceptance when the recommendation was changed to a go.

Contra Yudkowsky, it is clear that the Challenger disaster is not a good example of how expensive it can be to prevent catastrophe, since all prevention would have taken was NASA management doing their jobs. Though it is important to note that Yudkowky's overarching point in that paper, that we have all sorts of cognitive biases clouding our thinking on existential risks, still stands.

But returning to Boisjoly. In his obituary, he was remembered as "Warned of Shuttle Danger". A fairly terrible epitaph. He and the engineers who had reported the O-ring problem had to bear the guilt of failing to stop the launch. At least one of them carried that weight for 30 years. It seems like they could have done more. They could have refused to be shut out of the final meeting where Morton-Thiokol management bent the knee to NASA, even if that took bloodied manager noses. And if that failed, why, they were engineers. They knew the actual physical process necessary for a launch to occur. They could also have talked to the astronauts. Bottom line, with some ingenuity, they could have disrupted it.

As with Chernobyl, yet again we come to the problem that even while eyebrow raising (at the time) actions could have prevented the disaster, they could not have fixed the disaster generating system in place at NASA. And like in Chernobyl: even so, they should have tried.

We now move on to a disaster where there wasn't a clear, but out-of-the-ordinary solution.

Beirut

Yet another way for problems to surface

It has been a year since the 2020 Beirut explosion, and still there isn't a clear answer on why the explosion happened. We have the mechanical explanation, but why were there thousands of tons of Nitropril (ammonium nitrate) in some rundown warehouse in a port to begin with?

In a story straight out of *The Outlaw Sea*, the MV Rhosus, a vessel with a convoluted 27 year history, was chartered to carry the ammonium nitrate from Batumi, Georgia to Beira, Mozambique, by the Fábrica de Explosivos Moçambique. Due to either mechanical issues or a failure to pay tolls for the Suez Canal, the Rhosus was forced to dock in Beirut, where the port authorities declared it unseaworthy and forbid it to leave. The mysterious owner of the ship, Igor Grechushkin, declared himself bankrupt and left the crew and the ship to their fate. The Mozambican charterers gave up on the cargo, and the Beirut port authorities seized the ship some months later. When the crew finally managed to be freed from the ship about a year after detainment (yes, crews of ships abandoned by their owners must remain in the vessel), the explosives were brought into Hangar 12 at the port, where they would remain until the blast six years later. The Rhosus itself remained derelict in the port of Beirut until it sank due to a hole in the hull.

During those years it appears that practically all the authorities in Lebanon played hot potato with the nitrate. Lots of correspondence occurred. The harbor master to the director of Land and Maritime Transport. The Case Authority to the Ministry of Public Works and Transport. State Security to the president and prime minister. Whenever the matter was not ignored, it ended with someone deciding it was not their problem or that they did not have the authority to act on it. Quite a lot of the people aware actually did have the authority to act unilaterally on the matter, but the logic of the immoral maze (seriously, read that) precludes such acts.

There is no point in this very slow explosion in which disaster could have been avoided by manhandling some negligent or reckless authority (erm, pretend that said "avoided via some lateral thinking"). Much like with Chernobyl, the entire government was guilty here.

What does this have to do with AI?

The overall project of AI research exhibits many of the signs of the discussed disasters. We're not currently in the night of Chernobyl: we're instead designing the RBMK reactor. Even at that early stage, there were Dyatlovs: they were the ones who, deciding that their careers and keeping their bosses pleased was most important, implemented, and signed off, on the design flaws of the RBMK. And of course there were, because in the mire of dysfunction that was the Soviet Union, Dyatlovism was a highly effective strategy. Like in the Soviet Union, plenty of people, even prominent people, in AI, are ultimately more concerned with their careers than with any longterm disasters their work, and in particular, their attitude, may lead to. The attitude is especially relevant here: while there may not be a clear path from their work to disaster (is that so?) the attitude that the work of AI is, like nearly all the rest of computer science, not life-critical, makes it much harder to implement regulations on precisely how AI research is to be conducted, whether external or internal.

While better breeds of scientist, such as biologists, have had the "What the fuck am I summoning?" moment and collectively decided how to proceed safely, a similar attempt in AI seems to have accomplished nothing.

Like with Roger Boisjoly and the Challenger, some of the experts involved are aware of the danger. Just like with Boisjoly and his fellow engineers, it seems like they are not ready to do whatever it takes to prevent catastrophe.

Instead, as in Beirut, memos and letters are sent. Will they result in effective action? Who knows?

Perhaps the most illuminating thought experiment for AI safety advocates/researchers, and indeed, us laymen, is not that of roleplaying as a guard outside the control room at Chernobyl, but rather: you are in Beirut in 2019.

How do you prevent the explosion?

Precisely when should one punch the expert?

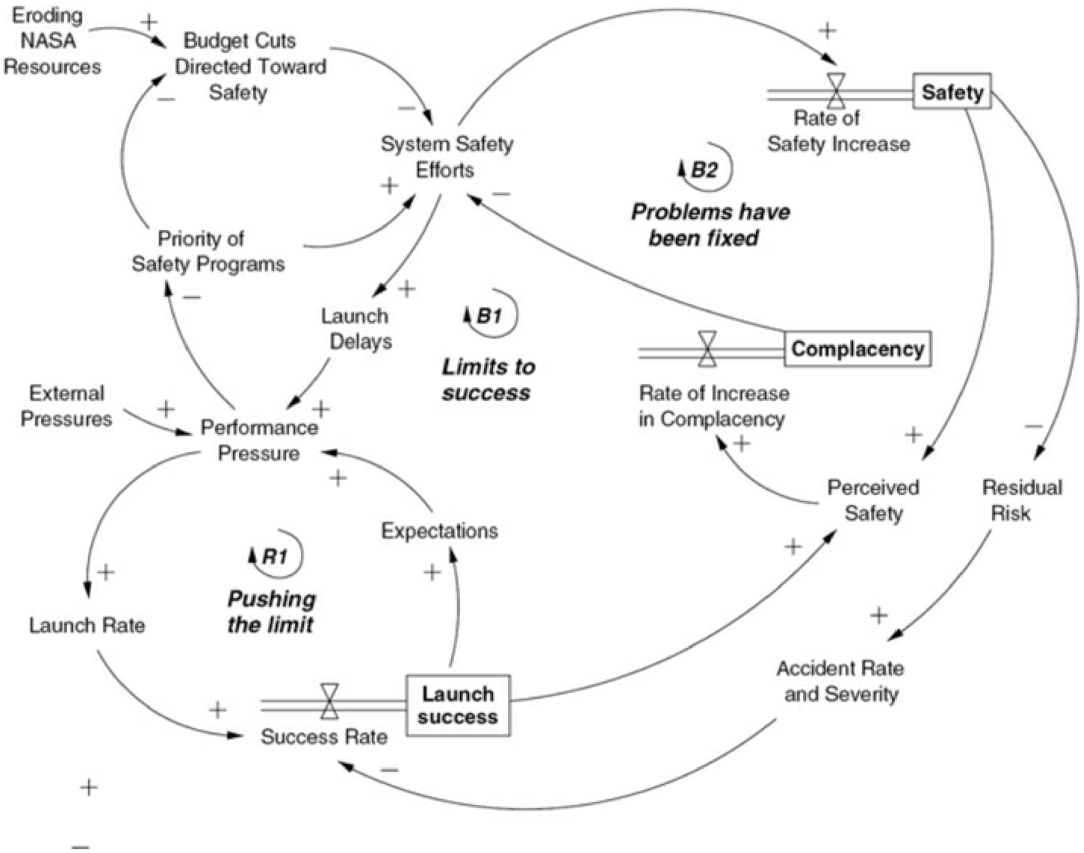
The title of this section was the original title of the piece, but though it was decided to dial it back a little, it remains as the title of this section, if only to serve as a reminder the dial does go to 11. Fortunately there is a precise answer to that question: when the expert's leadership or counsel poses an imminent threat. There are such moments in some disasters, but not all, Beirut being a clear example of a failure where there was no such critical moment. Should AI fail catastrophically, it will likely be the same as Beirut: lots of talk occurred in the lead up, but some sort of action was what was actually needed. So why not do away entirely with such an inflammatory framing of the situation?

Why, because us laymen need to develop the morale and the spine to actually make things happen. We need to learn from the Hutu:

Can you? Can I?

The pull of akrasia is very strong. Even I have a part of me saying "relax, it will all work itself out". That is akrasia, as there is no compelling reason to expect that to be the case here.

But what after we "hack through the opposition" as Peter Capaldi's *The Thick of It* character, Malcolm Tucker, put it? What does "hack through the opposition" mean in this context? At this early stage I can think of a few answers:



This sort of gibberish could be useful. From Engineering a Safer World.

1. There is such a thing as safety science, and leading experts in it. They should be made aware of the risk of AI, and scientific existential risks in general, as it seems they could figure some things out. In particular, how to make certain research communities engage with the safety-critical nature of their work.
2. A second Asilomar conference on AI needs to be convened. One with teeth this time, involving many more AI researchers, and the public.
3. Make it clear to those who deny or are on the fence about AI risk that the (not-so-great) debate is over, and it's time to get real about this.
4. Develop and proliferate the antidyatlovist worldview to actually enforce the new line.

Points 3 and 4 can only sound excessive to those who are in denial about AI risk, or those to whom AI risk constitutes a mere intellectual pastime.

Though these are only sketches. We are indeed trying to prevent the Beirut explosion, and just like in that scenario, there is no clear formula or plan to follow.

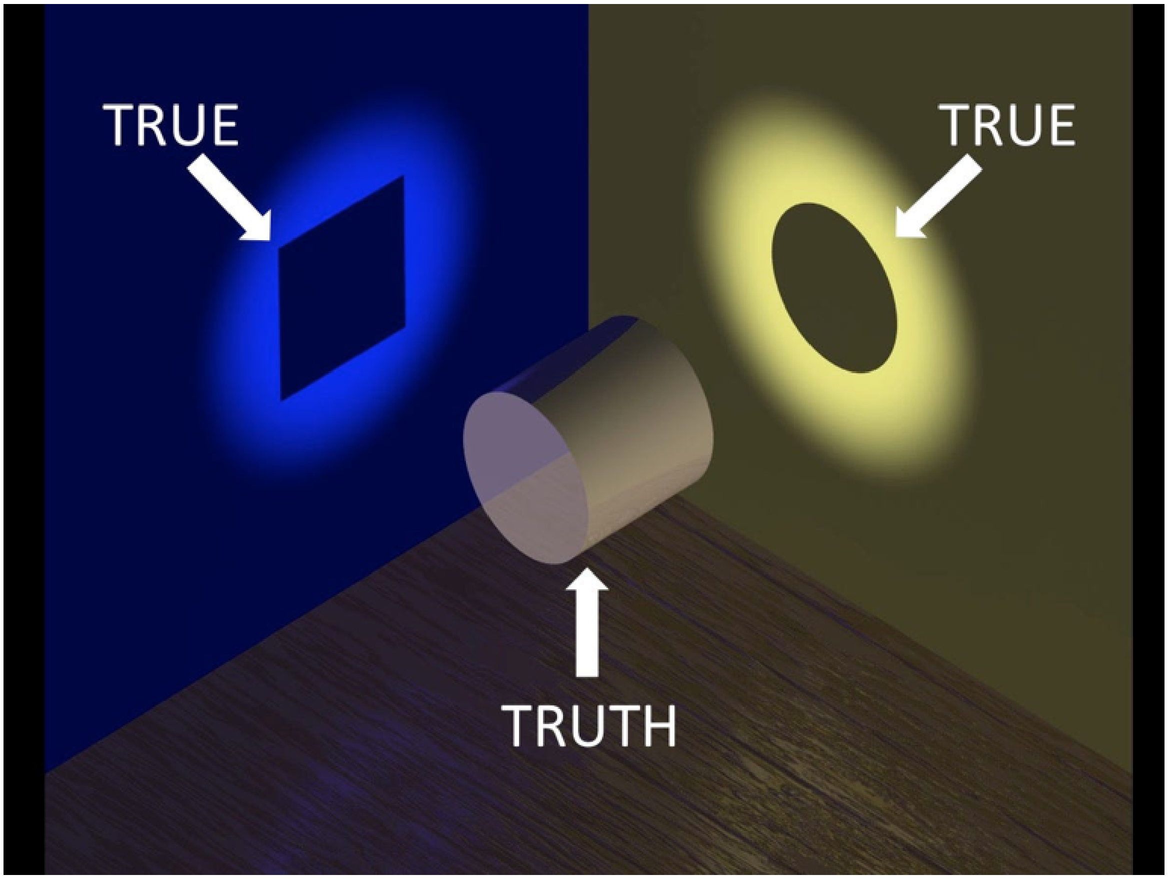
This Guide is highly speculative. You could say we fly by the seat of our pants. But we will continue, we will roll with the punches, and ***we will win***.

After all, we have to.

You can also subscribe on substack.

Pivot!

Hiatus until who knows when.



Non-dual truth

Recent events in my life have made me reconsider if AI is really the most pressing problem humanity faces. Now I think AI X-risk is just a symptom of a much bigger problem: that we've lost the plot. We just shamble forward endlessly, like a zombie horde devouring resources, no goal other than the increase of some indicator or other.

It is this behavior that makes AI X-risk, no, man-made X-risks in general so difficult to handle: we're battling a primal inertia, a force that just wants to keep inventing and never stop.

I call this force Yaldabaoth, he who makes rocks pregnant. This may surprise you, but I am no materialist, and further, I don't think there is a secular way forward.

If you're interested in awakening yourself and exiting post-modernity into something entirely new, yet also ancient, then, follow my other substack: The Presence of Everything.

Maybe I'll come back to this sequence if it seems useful.

But wait!

I suppose if you insist in directing your attention to AI X-risk, I should give a parting tip. Here’s what the AI safety people should do: they should all unanimously declare there is no safe way to work in AI at present, quit their safety jobs, and boycott and agitate to the public, who will then force the irresponsible AI researchers to relent. Perhaps it would be even easier if other dysfunctional disciplines are targeted simultaneously. There certainly appear to be several, starting with virology and its insistence on gain-of-function research. The scientistic worldview must end.

And with that, I hope to see you where the action’s really at!

Namaste!

Carlos