Re: I'm curious about forbidden ideas

1 message

Eliezer Yudkowsky <sentience@pobox.com> Tue, Jul 27, 2010 at 9:44 AM
To: ZJ <zjantolak@gmail.com>

DON'T THINK ABOUT THE AARDVARK

by

IF IT WAS A GOOD IDEA TO TALK ABOUT WHO THE AUTHOR WAS,

THEIR NAME WOULD BE WRITTEN HERE,

BUT IT'S NOT

NOT FOR PUBLICATION

COPY AND PASTE INTO EMAILS, DO NOT FORWARD

DO NOT MAKE DOWNLOADABLE

DO NOT SHARE THIS DOCUMENT WITH ANYONE

OR DISCUSS ITS CONTENTS WITH ANYONE

OR HINT AT ITS EXISTENCE TO ANYONE

UNLESS THEY ARE ALREADY HAVING AARDVARK PROBLEMS

OR THREATENING TO TALK ABOUT THE AARDVARK

DO NOT ENGAGE IN MYSTERIOUS TALK

ABOUT AARDVARKS IN FRONT OF OTHER PEOPLE

OR DO ANYTHING ELSE THAT WOULD MAKE THEM CURIOUS

THANK YOU FOR OBSERVING ALL SAFETY PRECAUTIONS

Q. What is this document about?

A. It's about the aardvark, and why you shouldn't think about it. Furthermore, so as to help other people not think about the aardvark, you should never talk about the aardvark with anyone.

Q. And what's the aardvark?

A. Not very fast on the uptake, are we?

Q. I see. So the aardvark is some sort of Langford Basilisk, or Snow Crash, or Riddle Theory, or weaponized joke a la Monty Python? A dangerous thought?

A. Sort of. The aardvark only happens if someone (1) thinks about the aardvark in (2) sufficient detail and (3) allows it to influence their actions in (4) a manner that is sufficiently predictable to be exploitable that (5) is not compensated for by some sort of counteracting influence. It's possible that (6) no one currently knows the sufficient detail, or even that (7) no human mind can know the sufficient detail; so (8) it would probably be very hard to think of the sufficient detail by accident while trying to think of something else; and (9) even if events 1-4 *did* happen, we could apply a compensating influence afterward that would cancel out the expected effect of thinking about the aardvark, which in turn would prevent the aardvark from happening. I'm telling you all of this information because it will help you to think less about the aardvark if you know that it's probably pretty hard to think about it successfully. The problem is that none of 6-9 is *certain*, and if all of events 1-5 actually did happen, it would be extremely bad for reasons that will go unnamed.

Q. ...how can you *possibly* expect me to take this seriously?

A. Yes! That's it! That's perfect! Just go on thinking that!

Q. No, I mean it. The notion of "dangerous thoughts" almost always relies on a supernatural model of reality - people who reason as if thought is ontologically fundamental and the universe has rules about thoughts, rather than thought being a phenomenon built on top of atoms which themselves obey the simple, mathematically regular, and local laws of physics. The laws of physics as we know them dictate that what goes on inside the skull stays inside the skull, unless it exits through ordinary means like sending nervous impulses through the spinal cord to muscles. I am therefore certain that you are wrong about the dangers of the aardvark.

A. Excellent.

Q. Which means there's no reason for me not to track down the real topic of this discussion and talk about it with people.

A. In that case, please rest assured that the aardvark is putatively embedded into a strictly reductionist universe that runs on the known laws of physics, and some extremely clever people who aspire to be very strict rational thinkers have decided that it cannot be ruled out, and the utility intervals involved are large enough to still be significant after being multiplied by a small probability. We should be extremely conservative when it comes to not thinking about the aardvark in sufficient detail.

Q. Sounds like Pascal's Wager to me.

A. Not everything that involves a small probability of a large utility is wrong merely because Pascal's original Wager was wrong. The problem with Pascal's Wager is that it involves taking a complicated possibility of extremely low prior probability and privileging it for consideration for no valid reason, while failing to consider any of the counterbalancing possibilities. The aardvark isn't being pulled out of thin air, it's something we have specific reason to worry about, and it's not automatically counterbalanced by anti-aardvarks.

Q. I tend to operate based on pattern recognition and perceptual categorization rather than detailed analysis and causal modeling, so this still "sounds like Pascal's Wager" and "sounds supernatural" to me. Therefore there are no real dangers

involved. Therefore it's okay to talk about it. QED.

A. *Also* the topic has already caused actual psychological damage to specific real people, so you can definitely hurt people by *talking* about the aardvark even if no such thing actually exists. This has *already happened*.

Q. How is that possible?

A. Do you have obsessive-compulsive disorder, or have you ever personally known anyone who leans toward the OCD side of the personality spectrum?

Q. No...

A. Okay. There are people in the world who are far more vulnerable than you to being told not to think about something because bad things will happen if they do. Not all of them have been diagnosed with clinical OCD, either; it's more of a personality spectrum. If you've never known anyone who's spent more than an entire year being tormented for hours at a time by thoughts of what they should've said to their last girlfriend to stop them from breaking up, thoughts which only come back stronger when they try to shove them away, to the point that they're experiencing difficulty functioning in everyday life - and these are people who *haven't* been diagnosed with OCD - then you don't realize how much very real, very painful damage you can do to someone whose personality tends toward that end of the spectrum by telling them about the aardvark.

Q. That all sounds a little implausible -

A. It's already happened to more than one person. Personally, I wouldn't have guessed it was probable either - but it *happened*. If my model of reality says that the actual observed outcome is improbable, this is a problem with my model, not a problem with reality. You may just not appreciate how much other people don't think the same way you do. Heck, I conceptualized the aardvark years ago and I didn't think it was much of a problem, because to me it was obvious that the equilibrium solution to the problem was that no rational agent would allow themselves to be influenced by thoughts of the aardvark, and even irrational agents like humans would be hard put to *accidentally* think about the aardvark in sufficient detail if we just started humming to ourselves whenever our thoughts threatened to go there -

Q. What about meta-aardvarks?

A. Rational agents ignore them, which prevents them from happening in the first place. You just don't go there; there's no positive incentive to do so.

Q. What about meta-meta-aardvarks -

A. Inverse Nike solution: You just don't. At the root of the whole proper class of ordinals, you just don't. There's no positive reason why a rational agent should ever go there in the first place, and so long as no one does, there won't be any aardvarks.

Q. I'm not sure that's how it works out. I see this possible objection -

A. I already thought of your objection, probably years ago, and it's wrong. Likewise if you think you see some obvious reason why there *can't* be aardvarks. That's wrong too.

Q. Could you be more specific?

A. Not without giving you information that could possibly contribute to thinking about the aardvark in sufficient detail. You'll just have to trust me.

Q. I'm engaging in a selective search for reasons why I should be allowed to know more, without a counterbalancing search for reasons why I shouldn't be, and I've hit upon the following argument: "Why should I trust you?"

A. If you know who the author of this document is, and you pretend that you find it implausible that the author knows more than you do, or you pretend that the silence likely reflects some kind of elaborate evil plot, then you are probably reasoning in a rather motivated fashion about the issue - demanding proof in a case where proof obviously cannot and should not be provided, and where proof is not necessary to choose based on expected utilities, because the demand for proof seems like a defense of your gut urge to talk about the subject or satisfy your curiosity. If you *don't* know and can't guess who the author of this document is, then you're probably pretty darned ignorant of the topic, no offense, but you hopefully have some kind of friend who you can trust.

Q. This all sounds increasingly absurd -

A. I'm fine with that.

Q. In fact, I begin to suspect you are crazy, or that this is an elaborate joke.

A. And the other reason not to talk about the aardvark is that it is unbelievably terrible PR. It just sounds completely crazy to anyone who doesn't know the basics of armadillo theory.

Q. "Armadillo theory."

A. You got a problem with that?

Q. You're giggling to yourself while you write this, aren't you.

A. It's been written in a deliberately silly style as a partial precaution against the damage it could do if it fell into the wrong hands. I.e., people would think it was an elaborate fictional joke essay. And it would be hard to make this essay sound deliberately silly if I did not, myself, possess a sense of humor. So, yes, giggling to myself.

Q. *Has* anyone actually talked about the aardvark?

A. When I originally thought of the aardvark, I said to myself, "Surely anyone smart enough to think of the aardvark would be smart enough not to talk about it in public." And this, I'm afraid, proved too optimistic. Apparently you actually *can* be that smart and yet still be that stupid. Someone ended up in a position where talking about the aardvark would let them - gasp! - *help win an argument*. And worse, the aardvark sounds like a clever idea, and there are people in this world who absolutely cannot bear the thought of being seen as less clever than they are. And who, for this reason, just cannot manage to shut up about an idea that will make them sound clever, no matter how damaging it is; they will find some excuse for why they need to talk about it anyway - exhibiting the usual pattern of "if I can find one single positive benefit, however remote, to be gained from talking about this idea, that means I should immediately run out and talk about it in public in this case where it will coincidentally help me win this argument I'm having". Fucking morons. I published my work on armadillo theory in part because I thought anyone clever enough to think up the aardvark in the first place would also see that the obvious equilibrium solution was not thinking about it, and have the *common sense* to never talk about it or hint at its existence, but no, clever fucking morons just can't manage to shut up.

Q. You sound a bit bitter.

A. I am, though I'm not so much bitter about damage being done by the actual aardvark, as bitter that clever morons are really that incapable of shutting up on this very elementary problem of shutting up, and capable of making something as harmless-seeming as armadillo theory into a dangerous idea that maybe shouldn't have been published. It has unpleasant implications for some other problems. On the plus side, there were some pretty funny conversations that ensued after the discussion of the aardvark was deleted, as the people who'd already seen the discussion, tried to convince the people who *hadn't* heard about the aardvark that they didn't *want* to hear about it. It went exactly the way you would expect it to: "No, trust me, you don't want to know." "But your telling me that makes me want to know MORE." "Look, it's not as cool as it sounds, okay? You can't do anything useful with it and it's not even all that important. If I tell you, you'll wish afterward that I hadn't." "Tell me! My mind is strong! I can handle it!" One important moral to take from all this is that unless someone has already heard about the aardvark, you don't tell them that there's an aardvark or that they're not allowed to know about it. And although I was laughing hysterically as I read the logs of the conversation, it was still sort of disheartening to find people behaving exactly like the stereotypical victims of an H. P. Lovecraft novel.

Q. People hunger for the forbidden knowledge, and the more forbidden it is, the more they hunger for it. They don't listen to the warnings, believing their minds invulnerable, unable to resist temptation, unable to stop poking and prodding at the unbearable itch of not knowing; and then afterward, they regret their curiosity horribly, but it's too late, all too late -

A. Yep, that's pretty much what happened only with less drama. Inevitably, they were unable to restrain their curiosity, and came up with cleverly stupid solutions like retrieving copies of the page from Google cache (now gone, thankfully). Most of them regretted their curiosity only very mildly, and many of them said "That's it? That's all?" But they did tend to agree that it was better not discussed further - once *they* knew about it, they went around trying to convince others not to ask the questions they themselves couldn't resist.

Q. And here I thought Lovecraft was fiction.

A. Well, all I can say is that the *human beings* in Lovecraft's novels turned out to have quite realistic reactions to the tempting prospect of forbidden knowledge, despite the fact that all the specific horror in Lovecraft is based on fictional supernatural entities. At least judging by the conversations that popped up when people had reason to take the prospect of something like a Langford Basilisk / Snow Crash / Riddle Theory / Weaponized Joke seriously.

Q. But you can't possibly expect people to suppress their curiosity!

A. Grownups can. If one of my trusted rationalist friends sent me a web address and told me that I must never visit it or something terrible would happen but they couldn't tell me what... then I would memorize the URL and never visit it. My significant other knows that I sometimes discuss private things with people, and when this occurs, goes elsewhere and

does not listen. Grownups can control themselves, both when it comes to keeping secrets - and the most important part of any secret is the fact that the secret exists, which is why you don't go around boasting to people that you know about this amazing secret - and they can also control themselves when it comes to the *resistible* urge to poke and prod at the lovely delicious forbidden Lovecraftian knowledge that a trusted friend asked them to leave alone. Just try to be a little more grown up.