Slate Star Codex

SELF-RECOMMENDING!

[REPOST] EPISTEMIC LEARNED HELPLESSNESS

[This is a slightly edited repost of an essay from my old LiveJournal]

A friend recently complained about how many people lack the basic skill of believing arguments. That is, if you have a valid argument for something, then you should accept the conclusion. Even if the conclusion is unpopular, or inconvenient, or you don't like it. He envisioned an art of rationality that would make people believe something *after it had been proven to them*.

And I nodded my head, because it sounded reasonable enough, and it wasn't until a few hours later that I thought about it again and went "Wait, no, that would be a terrible idea."

I don't think I'm overselling myself too much to expect that I could argue circles around the average uneducated person. Like I mean that on most topics, I could demolish their position and make them look like an idiot. Reduce them to some form of "Look, everything you say fits together and I can't explain why you're wrong, I just know you are!" Or, more plausibly, "Shut up I don't want to talk about this!"

And there are people who can argue circles around me. Maybe not on *every* topic, but on topics where they are experts and have spent their whole lives honing their arguments. When I was young I used to read pseudohistory books; Immanuel Velikovsky's [Ages in Chaos](#) is a good example of the best this genre has to offer. I read it and it seemed so obviously correct, so perfect, that I could barely bring myself to bother to search out rebuttals.

And then I read the rebuttals, and they were so obviously correct, so devastating, that I couldn't believe I had ever been so dumb as to believe Velikovsky.

And then I read the rebuttals to the rebuttals, and they were so obviously correct that I felt silly for ever doubting.

And so on for several more iterations, until the labyrinth of doubt seemed inescapable. What finally broke me out wasn't so much the lucidity of the consensus view so much as starting to sample different crackpots. Some were almost as bright and rhetorically gifted as Velikovsky, all presented insurmountable evidence for their theories, and all had mutually exclusive ideas. After all, Noah's Flood couldn't have been a cultural memory both of the fall of Atlantis *and* of a change in the Earth's orbit, let alone of a lost Ice Age civilization or of megatsunamis from a meteor strike. So given that at least some of those arguments are wrong and all seemed practically proven, I am obviously just gullible in the field of ancient history. Given a total lack of independent intellectual steering power and no desire to spend thirty years building an independent knowledge base of Near Eastern history, I choose to just accept the ideas of the prestigious people with professorships in Archaeology, rather

than those of the universally reviled crackpots who write books about [Venus being a comet](#).

You could consider this a form of epistemic learned helplessness, where I know any attempt to evaluate the arguments is just going to be a bad idea so I don't even try. If you have a good argument that the Early Bronze Age worked completely differently from the way mainstream historians believe, I just don't want to hear about it. If you insist on telling me anyway, I will nod, say that your argument makes complete sense, and then totally refuse to change my mind or admit even the slightest possibility that you might be right.

(This is the correct Bayesian action: if I know that a false argument sounds just as convincing as a true argument, argument convincingness provides no evidence either way. I should ignore it and stick with my prior.)

I consider myself lucky in that my epistemic learned helplessness is circumscribed; there are still cases where I'll trust the evidence of my own reason. In fact, I trust it in most cases other than infamously deceptive arguments in fields I know little about. But I think the average uneducated person doesn't and shouldn't. Anyone anywhere – politicians, scammy businessmen, smooth-talking romantic partners – would be able to argue them into anything. And so they take the obvious and correct defensive maneuver – they will never let anyone convince them of any belief that sounds "weird".

(and remember that, if you grow up in the right circles, beliefs along the lines of "astrology doesn't work" sound "weird".)

This is starting to resemble ideas like compartmentalization and [taking ideas seriously](). The only difference between their presentation and mine is that I'm saying that for 99% of people, 99% of the time, taking ideas seriously is the wrong strategy. Or, at the very least, it should be the last skill you learn, after you've learned every other skill that allows you to know which ideas are or are not correct.

The people I know who are best at taking ideas seriously are those who are smartest and most rational. I think people are working off a model where these co-occur because you need to be very clever to resist your natural and detrimental tendency not to take ideas seriously. But I think they might instead co-occur because you have to be really smart in order for taking ideas seriously not to be immediately disastrous. You have to be really smart not to have been talked into enough terrible arguments to develop epistemic learned helplessness.

Even the smartest people I know have a commendable tendency not to take certain ideas seriously. Bostrom's [simulation argument](), the [anthropic doomsday argument](), [Pascal's Mugging]() – I've never heard anyone give a coherent argument against any of these, but I've also never met anyone who fully accepts them and lives life according to their implications.

A friend tells me of a guy who once accepted fundamentalist religion because of Pascal's Wager. I will provisionally admit that this person "takes ideas seriously". Everyone else gets partial credit, at best.

Which isn't to say that some people don't do better than others.

Terrorists seem pretty good in this respect. People used to talk about how terrorists must be very poor and uneducated to fall for militant Islam, and then someone did a study and found that they were disproportionately well-off, college educated people (many were engineers). I've heard a few good arguments in this direction before, things like how engineering trains you to have a very black-and-white right-or-wrong view of the world based on a few simple formulae, and this meshes with fundamentalism better than it meshes with subtle liberal religious messages.

But to these I'd add that a sufficiently smart engineer has never been burned by arguments above his skill level before, has never had any reason to develop epistemic learned helplessness. If Osama comes up to him with a really good argument for terrorism, he thinks "Oh, there's a good argument for terrorism. I guess I should become a terrorist," as opposed to "Arguments? You can prove anything with arguments. I'll just stay right here and not blow myself up."

Responsible doctors are at the other end of the spectrum from terrorists here. I once heard someone rail against how doctors totally ignored all the latest and most exciting medical studies. The same person, practically in the same breath, then railed against how [50% to 90% of medical studies are wrong](). These two observations are not unrelated. Not only are there so many terrible studies, but pseudomedicine (not the stupid homeopathy type, but [the type that]() links everything to some obscure chemical on an out-of-the-way metabolic pathway) has, for me, proven much like pseudohistory – unless I am an expert in that particular subsubfield of medicine, it can sound very convincing even when it's very wrong.

The medical establishment offers a shiny tempting solution. First, a total unwillingness to trust anything, no matter how plausible it sounds, until it's gone through an endless cycle of studies and meta-analyses. Second, a bunch of Institutes and Collaborations dedicated to filtering through all these studies and analyses and telling you what lessons you should draw from them.

I'm glad that some people never develop epistemic learned helplessness, or develop only a limited amount of it, or only in certain domains. It seems to me that although these people are more likely to become terrorists or Velikovskians or homeopaths, they're also the only people who can figure out if something basic and unquestionable is wrong, and make this possibility well-known enough that normal people start becoming willing to consider it.

But I'm also glad epistemic learned helplessness exists. It seems like a pretty useful social safety valve most of the time.

388 RESPONSES TO *[REPOST] EPISTEMIC LEARNED HELPLESSNESS*

theodidactus *says:*

June 3, 2019 at 5:37 pm

Law student here: I'm coming down off a large project that involved researching the history of eugenics, particularly the legal aspects. Virtually everyone involved from the supreme court on down had VERY good, VERY convincing arguments for compulsory sterilization (both legal and scientific).

I enjoy pointing out that in Buck v. Bell, the only time the supreme court has touched this issue (and upheld a compulsory sterilization law) the lone dissent was not the uber-progressive Louis Brandeis, or the chimeric but utterly brilliant Oliver Wendall Holmes, or most of the court's conservative wing that later opposed the New Deal almost whole cloth... it was the curmudgeonly catholic, Pierce Butler...who was just about the most unlikeable guy to a modern legal analyst. He was wrong about EVERYTHING...except this.

Maybe he just got lucky. His dissent was silent. I can't say whether it was his catholicism, his curmugeonlyness, or something else, that lead him to what I must assume is the "correct" conclusion that the state cannot compel sterilization using the same powers it uses to, say, compel vaccination..but it intrigues me that somehow, he was the ONLY ONE that got this right.

HideHide

Doctor Locketopus *says:*

June 3, 2019 at 5:55 pm

Hmm... according to Wikipedia (yeah, I know) Butler was also the lone dissenter in Palko v. Connecticut, which decided 8-1 that the Fifth Amendment prohibition of double jeopardy did not apply to the states (Butler thought it clearly did apply, based on the 14th Amendment). Palko was executed. Butler's position was later vindicated when Palko v. Connecticut was overturned by Benton v. Maryland.

He did oppose many of Roosevelt's New Deal power-grab laws, rightly in my opinion.

Hide ↑Hide ↑

Clutzy *says:*

June 3, 2019 at 9:11 pm

Lawyer here. Just going to disagree with the Holmes was brilliant thing. I think he was trendy and fairly pedestrian outside of that.

Hide ↑Hide ↑

theodidactus *says:*

June 4, 2019 at 10:53 am

I think Holmes should get major points simply for being such an exceptionally good writer/explainer. Surely you can give him that, right?

Hide ↑Hide ↑

Clutzy *says:*

June 4, 2019 at 1:12 pm

He was a good writer. As an explainer I'd say he is fairly deficient. As least when it comes to explaining why he came to his trendy legal conclusions.

Hide ↑Hide ↑

Steve Sailer *says:*

June 4, 2019 at 1:49 am

If you look at sterilization laws by state, Catholic states were less likely to fall for the eugenics fad. Southern Baptist states were in the middle, and mainline progressive Protestant states trending toward secularism tended to be the most enthusiastic for eugenics.

John Glad's book "Jewish Eugenics" points out that Jews tended to be fairly enthusiastic for eugenics until the late 1960s.

So the votes of Holmes (who was pretty much of a nihilist Protestant) and Brandeis were not out of line with their demographics.

Hide ↑Hide ↑

theodidactus *says:*

June 4, 2019 at 6:43 am

Absolutely, I've pointed that out before, largely in the context of saying that my own views tend to eerily conform to Catholicism even though I renounced the religion decades ago. Maybe we're just products of what we learn in the first 10 or so sessions of sunday school.

Hide ↑Hide ↑

pqjk2 *says:*
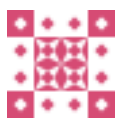
June 4, 2019 at 9:30 am

username checks out

Hide ↑Hide ↑

MereComments *says:*

June 4, 2019 at 9:42 am

Or, maybe your views tend eerily towards Catholicism because Catholicism reflects the truth of the world. 😄

Hide ↑Hide ↑

theodidactus *says:*

June 4, 2019 at 10:58 am

Well if that's true it's weird I absorbed all the stuff on the periphery, but none of the stuff at the hub (to my memory I believed in god about as long as I believed in Santa Claus).

Here's the fun one I bring up a lot. I'm atheist, by my understanding of the term. I hang out with a lot of other atheists...but most of them are ex-protestant. Their whole approach strikes me as VERY different from the other ex-catholic atheists I know. Not to knock them (or all the wonderful ex-protestant atheists out here) but I personally feel that ex-catholic atheists are much more amenable to chesterton's-fence-type arguments, especially when they involve tolerating people with views and opinions very different from your own. (One can obviously see how this relates to the Eugenics issue).

I also notice Ex-catholic atheists in my experience anyway have a much stronger attraction to supernatural/pseudoscientific concepts, even if they don't believe them. (Hence the screenname). When I start talking about alchemy or scientology most atheists are like "well, that's some bullshit. Next topic." but I notice ex-catholic atheists seem to be more likely to be like "oooh cool!"

Hide ↑Hide ↑

Steve Sailer *says:*

G.K. Chesterton, a convert to Catholicism, published a 1922 book entitled "Eugenics and Other Evils:"

https://en.wikisource.org/wiki/Eugenics_and_other_Evils

Chesterton's frenemy G.K. Shaw was a famous promoter of eugenics, as was H.G. Wells. There was a dinner for Francis Galton in 1904 in which Shaw and Wells made over-the-top extremist pro-eugenics speeches. Galton made a few cryptic remarks at the end, which I interpret as him rolling his eyes at how far Shaw and Wells were taking his idea (although I might be wrong about that).

Hide ↑Hide ↑

Murphy *says:*

June 4, 2019 at 3:41 am

This is something I find with a number of things.

When it comes to issues where a huge fraction of the population held or hold a position…. normally there's actually a pretty solid philosophical foundation.

I think I mentioned in the last thread that at some point I want to try to create a flow chart of all the reasonably coherent/consistent positions on abortion and their philosophical underpinnings since that's one area I used to enjoy digging into.

I'm still pro-choice but that's because I still have the same precepts… but I can see very well that a few changes of precepts would lead to a different conclusion.

But as a result I find it almost distressing to see almost any /r/politics type discussion on the subject where people genuinely seem to believe that their opponents could only ever reach their position from a basis of "we hate women"

People seem almost afraid of letting anyone hear the coherent and logical versions of their opponents positions.

But the same goes for lots of things. Eugenics is one such. Story book eugenicists

stand around kackling about wanting to advance evil. Real world eugenicists were able to point to children suffering horribly and say that that was a bad thing.

To anyone familiar with animal husbandry eugenics would seem to fall out as a natural consequence. You can have a healthy herd in the wild by letting predators cull the sickly and weak or in captivity you have to do something else if you don't want to see a herd of very sickly animals. The call for someone to manage the group's reproduction would naturally fall out of the same kind of sentiments that call for educating the poor ,feeding the hungry and protecting the shared commons.

If my precepts didn't instead lead me to transhumanism and the general belief that it'll all be a bit irrelevant as we get better at manipulating biology directly it would seem very logical to me. But since I don't believe it will be an issue for enough generations to really matter it does not.

Hide ↑Hide ↑

theodidactus *says:*

June 4, 2019 at 6:28 am

Minnesota's sterilization regime is the best example of "well meaning" eugenics I've found. It grew directly out of its child protection and welfare system, and all the primary sources I've encountered lead me to believe two of its three progenitors were extremely concerned well meaning people who cared a great deal about the people they were sterilizing. The third fellow, Dight (wiki him), was a bit of a fanatic and famously wrote fanmail to Hitler.

Hide ↑Hide ↑

Jiro *says:*

June 4, 2019 at 7:34 am

> *When it comes to issues where a huge fraction of the population held or hold a position.... normally there's actually a pretty solid philosophical foundation.*

This fails when it comes to memes (in the pre-Internet meme sense). A huge fraction of the population believes in transsubstantiation because Catholicism succeeded in killing and converting enough people and getting them to teach itself to their children.

There's a philosophical foundation, of course, but the merits of that philosophical foundation are irrelevant (and the fact that non-Catholics don't accept it seems to indicate it's worthless).

The original Mr. X *says:*

June 4, 2019 at 8:03 am

> *This fails when it comes to memes (in the pre-Internet meme sense). A huge fraction of the population believes in transsubstantiation because Catholicism succeeded in killing and converting enough people and getting them to teach itself to their children.*

"Catholicism" did no such thing, because "Catholicism" is an abstraction and hence not an agent. What actually happened is that people were persuaded by other people to convert and then taught their new beliefs to their children.

> *There's a philosophical foundation, of course, but the merits of that philosophical foundation are irrelevant (and the fact that non-Catholics don't accept it seems to indicate it's worthless).*

It indicates nothing of the sort; all it indicates is that people who accept the philosophical foundations of Catholicism generally become Catholics. Which, like, duh, what did you expect?

Jiro *says:*

June 5, 2019 at 10:49 am

> *all it indicates is that people who accept the philosophical foundations of Catholicism generally become Catholics.*

To a very close approximation, nobody becomes Catholics

because they accept the philosophical foundations of Catholicism. They become Catholics because they were raised Catholic, or perhaps to marry into a Catholic family, or some similar reason.

You don't see a population of people not raised Catholic who come to say "transsubstantiation makes sense, and so I have to become Catholic". If there was any substance to transsubstantiation, you would expect this to happen.

Hide ↑ Hide ↑

Mary *says:*

June 5, 2019 at 5:50 pm

Perhaps you are not looking in the right place. I have certainly heard of people who became Catholics because they became convinced of transsubstantiation.

Hide ↑ Hide ↑

Murphy *says:*

June 4, 2019 at 8:17 am

philosophical foundation vs factual foundation.

If you believe in souls and life after death (something a lot of people believe in because people tend to really really want to believe there's something) then various things fall out of that.

that we have no proof of the existence of such puts an awful big hole in the factual foundation but not the philosophical.

It's like a mathematical proof that starts with "assuming XYZ to be true", it still a perfectly good proof even if XYZ turns out to not be true in the real world.

Hide ↑ Hide ↑

Mary *says:*

Do you have proof that we do not have proof of the existence of such? After all, a lot of people really really want to believe there's no such proof.

Hide ↑Hide ↑

notpeerreviewed *says:*

I'm not sure cackling was involved, but most real-world eugenicists do not come off well to modern readers. They justified eugenics almost exclusively in collectivist terms – national strength, progress of the Caucasian race, and stuff like that. Do you have an example of a prominent eugenicist who relied heavily on arguments about suffering children?

Hide ↑Hide ↑

Deiseach *says:*

I'm not sure cackling was involved, but most real-world eugenicists do not come off well to modern readers.

An example from 1895, from a [story] entitled "The S.S." by M.P. Shiel – it's **very** florid prose, about a reclusive detective in the vein of Poe's Chevalier Dupin who investigates a rash of suicides and murders, and discovers a secret "Society of Sparta" which takes upon itself the eugenic duty to weed out the unfit (so it painlessly poisons both rich and poor who are producing hordes of sickly children, and these mysterious deaths inflame the susceptible into copy-cat suicides).

The protagonist, if not approving of how they're going about it, does approve of the principles behind it, and has a long speech about the unintended consequences of modern medicine allowing the sickly and unfit to live (I've pruned it down as much as I could):

> *And now let me apply these facts to the Europe of our*
> *own time. We no longer have world-serious war — but in*

its place we have a scourge, the effect of which on the modern state is precisely the same *as the effect of war on the ancient, only, — in the end, — far more destructive, far more subtle, sure, horrible, disgusting. The name of this pestilence is Medical Science. Yes, it is most true, shudder — shudder — as you will! Man's best friend turns to an asp in his bosom to sting him to the basest of deaths. The devastating growth of medical, and especially surgical, science — that, if you like, for us all, is "the question of the hour!" And what a question! of what surpassing importance, in the presence of which all other "questions" whatever dwindle into mere academic triviality. For just as the ancient State was wounded to the heart through the death of her healthy sons in the field, just so slowly, just so silently, is the modern receiving deadly hurt by the botching and tinkering of her unhealthy children. The net result is in each case the same — the altered ratio of the total amount of reproductive health to the total amount of reproductive disease. …We are at this very time, if I mistake not, on the verge of new insights which will enable man to laugh at disease — laugh at it in the sense of over-ruling its natural tendency to produce death, not by any means in the sense of destroying its ever-expanding* existence. *Do you know that at this moment your hospitals are crammed with beings in human likeness suffering from a thousand obscure and subtly-ineradicable ills, all of whom, if left alone, would die almost at once, but ninety in the hundred of whom will, as it is, be sent forth "cured," like missionaries of hell, and the horrent shapes of Night and Acheron, to mingle in the pure river of humanity the poison-taint of their protean vileness? Do you know that in your schools one-quarter of the children are already purblind? Have you gauged the importance of your tremendous consumption of quack catholicons, of the fortunes derived from their sale, of the spread of modern nervous disorders, of toothless youth*

and thrice loathsome age among the helot-classes? Do you know that in the course of my late journey to London, I walked from Piccadilly Circus to Hyde Park Corner, during which time I observed some five hundred people, of whom twenty-seven only were perfectly healthy, well-formed men, and eighteen healthy, beautiful women? ...Less death, more disease — that is the sad, the unnatural record; children especially — so sensitive to the physician's art — living on by hundreds of thousands, bearing within them the germs of wide-spreading sorrow, who in former times would have died. And if you consider that the proper function of the doctor is the strictly limited one of curing the curable, rather than of self-gloriously perpetuating the incurable, you may find it difficult to give a quite rational answer to this simple question: why?

So any of you who had to wear glasses as kids ('one-quarter of the children already purblind in the schools'), and are maybe now all grown up and married with kids of your own, how very dare you shapes of Night and Acheron mingle the poison-taint of your protean vileness in the pure river of humanity? 🙂

Hide ↑Hide ↑

theodidactus *says:*

June 4, 2019 at 4:08 pm

Mingling of poison taints is what humanity is all about!

Hide ↑Hide ↑

vV_Vv *says:*

June 4, 2019 at 4:49 am

It's not obvious to me that eugenics is wrong.

Most arguments I've heard about it boil down to "Hitlerdidit!!!" or "it had a disparate impact on blacks/gypsies/[other underperforming minority]".

Slightly more sophisticated arguments invoke bodily autonomy, but these don't really work for mentally incompetent people who aren't generally considered to have bodily autonomy and are in fact [usually prohibited from having sex](), and certainly don't apply to voluntary eugenics practices.

Does anybody have a compelling argument against eugenics?

Murphy *says:*

June 4, 2019 at 5:16 am

You mean in the broadest sense that includes parents making choices about their own offspring or in the narrow sense of the state controlling reproduction to achieve some kind of end goal related to some desired traits or health of the populous?

Inpatient mental health care and an institutions desire to avoid lawsuits is not quite representative of the general case. For people with mental problems living in the community and people with even quite significant cognitive impairments the default tends to be not to restrict them from sex.

Often courts are extremely hesitant to entirely restrict someone from sexual activity because it's such a significant source of happiness for many people. The bar tends to be pretty low: understanding the basics of sex and some sexual health stuff.

But even then, the basis of that is more a restriction **on others** than on the individual in question. That a mentally able person would be taking advantage of them rather than there being anything inherently wrong with them wanting to have sex themselves.

I don't believe your link serves as a precedent that leads to negation of a right to bodily autonomy or bodily integrity any more than parents being allowed make their kids do chores is a good precedent for slavery.

What you'd find compelling kind of depends on your own precepts.

If you have a strong inclination towards "the good of the many"/group over individual / communitarianism then it's easier to argue for eugenics. It sucks for people who want kids but who get disallowed from having them but it also sucks for the people who lost the lottery and ended up divided up to deal with the organ shortage at the hospitals.

If you have a strong inclination towards individualism then involuntary eugenics is kinda gonna fly in the face of a lot of your precepts.

Hide ↑Hide ↑

vV_Vv *says:*

June 5, 2019 at 6:33 am

> *Often courts are extremely hesitant to entirely restrict someone from sexual activity because it's such a significant source of happiness for many people. The bar tends to be pretty low: understanding the basics of sex and some sexual health stuff.*

Doesn't this conflict with the principle that legitimate sexual activity requires informed consent? If somebody isn't considered able to make their own decision about, e.g. what eat for dinner, how can they possibly be able to make their own decision about something that can result in pregnancy or STD?

> *If you have a strong inclination towards "the good of the many"/group over individual / communitarianism then it's easier to argue for eugenics.*

You can argue for eugenics from an individualist point of view by pointing to the positive or negative externalities produced by people with certain traits. Welfare for the disabled and law enforcement to deal with criminals, all payed for with taxpayer money, are the most obvious examples.

There is also the argument that we may have an obligation to prevent people who would live lives full of suffering from being born, if possible. I'm not sure if it is an individualist or collectivist

argument.

Aapje *says:*

June 5, 2019 at 6:41 am

@Murphy

> *Often courts are extremely hesitant to entirely restrict someone from sexual activity*

This is not actually possible without putting people in prison or using chastity devices, both of which have consequences beyond impacting sexual activity.

Perhaps you meant: forced birth control (permanent or temporary)?

PDV *says:*

June 10, 2019 at 8:15 pm

People in the hospital or outpatient mental healthcare facilities can easily be, and routinely are, prohibited from engaging in sexual activity.

Unless you count masturbation as sexual activity, which is a valid though silly way to draw your category boundaries; in that case they can easily be, and routinely are, prohibited from non-masturbation sexual activity.

theodidactus *says:*

June 4, 2019 at 6:35 am

The "right" to reproduce (let's call it a right) is hard to appreciate for a lot of people, and I might include myself in that category from time to time. Plenty of people get along just fine without offspring, and have no intention of having them. They find meaning in other areas of life.

But the same is true of what we say and believe. Plenty of people get by just fine without burning flags or drawing pictures of mohammed or watching really odd porn. Why do you need to go and do something crazy like that? It does clear damage to the social fabric, causes obvious distress to a certain category of people, and communicates nothing of real importance.

We recognize in the right to speak freely a good beyond even the marketplace of ideas…it is not just for the benefit of society that we let you watch what you want and say what you want…we do it because it's a really really important part of being a person. So is having the ability to reproduce (even if you never use it).

As another potential argument: Scott's post above, and some of his others (See his review of seeing like a state, eichmann in jerusalem) point to the social benefit of having a large, diverse population which includes several varieties of maniac, lazy good for nothing, idiot, coward, and anxious mess. Monocultures often suck. A forest is much more susceptible to burning down, or succumbing to diseases, if you fill it with a bunch of trees that all have the same desirable properties. A society is much more susceptible to all kinds of evil if it's full of a bunch of optimized people that all have the same responses to everything.

Hide ↑Hide ↑

Randy M *says:*

June 4, 2019 at 7:02 am

> *But the same is true of what we say and believe.*

I was going to make this comparison from another angle. Eugenics is like censorship–it's not that everyone's contribution is equally positive–some people are not even net positive. But we don't trust the state to be wise or disinterested enough to discriminate between them.

Especially when the ability to reproduce is so valued, and thus leverage for a corrupt state.

theodidactus *says:*

June 4, 2019 at 7:27 am

That's a good point I hadn't really considered before. My argument is almost always self actualization mixed with "everyone from the top down is bound to get this wrong" but it makes perfect sense to also say, as you did, "a corrupt government could easily misuse this power"

Hide ↑Hide ↑

Jiro *says:*

June 4, 2019 at 7:37 am

It's not just misuse. Governments can have bad judgment, and people who make rules about things that don't affect themselves can have exceptionally bad judgment about them.

Hide ↑Hide ↑

eyeballfrog *says:*

June 4, 2019 at 11:27 am

> *Plenty of people get along just fine without offspring, and have no intention of having them. They find meaning in other areas of life. But the same is true of what we say and believe. Plenty of people get by just fine without burning flags or drawing pictures of mohammed or watching really odd porn. Why do you need to go and do something crazy like that?*

This is a weird-sounding argument. Burning flags is something a tiny minority wants, but having children is something all *but* a tiny minority wants.

Hide ↑Hide ↑

theodidactus *says:*

June 4, 2019 at 11:33 am

Burning flags itself is a weird thing only a tiny minority wants, but self-expression is a thing a huge number of people want. So if we want a world where we can throw our arms wide and say "look, self-expression, isn't it wonderful!" We need to put up with some flag burning (or other suitably nasty and completely irrational expression you hate).

Similarly, having what looks, in your opinion, to be screwed up kids is a weird thing only a tiny minority of people want, but the ability to reproduce is a thing a huge number of people want. If we want a world where we can throw our arms wide and say "The ability to reproduce! Isn't it wonderful!" we have to put up with some stuff that is going, in our eyes, to look like bad reproductive decisionmaking.

Hide ↑Hide ↑

vV_Vv *says:*

June 5, 2019 at 6:14 am

> *But the same is true of what we say and believe. Plenty of people get by just fine without burning flags or drawing pictures of mohammed or watching really odd porn. Why do you need to go and do something crazy like that? It does clear damage to the social fabric, causes obvious distress to a certain category of people, and communicates nothing of real importance.*

I'd say the negative externalities of free speech are much less than the negative exteralities of producing certain kind of people: you can always ignore speech you don't like, while you still have to pay taxes to fund welfare for the disabled, the police, courts and prisons for the criminals, and so on.

> *As another potential argument: Scott's post above, and some of his others (See his review of seeing like a state,*

> *eichmann in jerusalem) point to the social benefit of*
> *having a large, diverse population which includes*
> *several varieties of maniac, lazy good for nothing, idiot,*
> *coward, and anxious mess. Monocultures often suck.*

Isn't this a fully general argument against, e.g. having laws?

Hide ↑Hide ↑

HeelBearCub *says:*

June 5, 2019 at 7:25 am

Here is someone who doesn't believe in the power of culture …

[slight snark, but I feel it is a legitimate point]

Hide ↑Hide ↑

notpeerreviewed *says:*

June 4, 2019 at 12:38 pm

I don't think there's anything wrong with "voluntary eugenics", e.g. you think your genes are good for society so you try to reproduce a lot; or less commonly, you think your genes are bad, so you don't reproduce. Historically, however, eugenicists didn't have much luck getting people to behave that way voluntarily, so most real-world eugenics policies were nonvoluntary.

Hide ↑Hide ↑

notpeerreviewed *says:*

June 4, 2019 at 12:46 pm

It's also worth noting that most of the actual historical eugenicists were extremely racist – not "product of their time" racist, but "writes literal fan letters to Hitler" racist. I mention this not to tar the idea by association, but to point out that eugenics reached the peak of its popularity when it was focused on demonizing an outgroup. The idea of altering one's *own* reproductive choices for the good of society has been tried and found to be a hard sell.

DavidFriedman *says:*

June 4, 2019 at 2:59 pm

> It's also worth noting that most of the actual historical eugenicists were extremely racist – not "product of their time" racist, but "writes literal fan letters to Hitler" racist.

Churchill? Shaw? Keynes?

theodidactus *says:*

June 4, 2019 at 3:59 pm

Dight

http://www2.mnhs.org/library/findaids/P1628/pdfa/P1628-00001.pdf

vV_Vv *says:*

June 5, 2019 at 6:03 am

So the US Supreme Court was staffed by Hitler fanboys?

emiliobumachar *says:*

June 5, 2019 at 2:32 am

Subsidizing and advertising contraceptives is a standard liberal value, and does drastically reduce the reproduction of the poor. It's just not called eugenics.

vV_Vv *says:*

June 5, 2019 at 6:05 am

> I don't think there's anything wrong with "voluntary

> *eugenics", e.g. you think your genes are good for society so you try to reproduce a lot; or less commonly, you think your genes are bad, so you don't reproduce.*

I think that "voluntary eugenics" usually means that the state will incentivize or disincentivize people with certain traits to reproduce more or less, usually by paying them or by some mean other than coercive power.

Hide ↑Hide ↑

Galle *says:*

June 4, 2019 at 12:56 pm

The central meaning of the word "eugenics" is "the killing, forced sterilization, and forced copulation of people in order to promote genetic traits that the state considers desirable." That may not be the literal definition of the word, but it's what most people think when you say it.

This sort of eugenics is obviously wrong for two reasons:

1. Killing, forced sterilization, and forced copulation are all pretty evil things to do.
2. The state does not have a great track record when it comes to correctly identifying what genetic traits to promote.

When you avoid mentioning the word "eugenics" and stick only to suggesting voluntary eugenic practices for eliminating genetic traits that everyone agrees are bad (like hereditary diseases) support for the practice is quite a bit higher. It's just when you start associating yourself with early twentieth century genocides that people find you questionable.

Hide ↑Hide ↑

notpeerreviewed *says:*

June 4, 2019 at 3:14 pm

I think you're mostly correct, but I'd like to add a rider: The historical eugenics movement promoted both "negative eugenics" and "positive eugenics" and I think. For the most part, what you're

talking about is "negative eugenics", and it's evil for the reasons you mention.

Negative eugenics policies were definitely more widely known, but positive eugenics is mentioned in history books often enough that I think it's still fairly central. Positive eugenics would technically include "forced copulation", but for the most part it involved voluntary policies that were quite a bit less objectionable, if maybe still kinda creepy. The biggest problem for positive eugenicists is that they rarely managed to get anyone to do it; it's pretty hard to rope people into voluntarily basing their childbearing decisions on the needs of the state. See, for example, the Repository for Germinal Choice and the difficulties they had getting volunteers.

Hide ↑Hide ↑

theodidactus *says:*

June 4, 2019 at 4:04 pm

Anyone read Doc EE Smith's "Lensman" Series? It's badly written but the progenitor of a ton of science fiction. A powerful alien race has been shaping the bloodline of the human race for eons. The final generation nears completion, and it becomes painfully obvious to the male and female protagonist that like, half the galaxy is in on a vast conspiracy to get them to hook up.

Hide ↑Hide ↑

BlindKungFuMaster *says:*

June 4, 2019 at 5:10 am

In the long run eugenics is necessary or we are going to hit a very low mutation-selection balance. Non-coercive eugenics seems to be obviously good (and of course is practiced in every partner choice). And eugenics based on gene modification is only problematic until it just works. Then it is as morally imperative as vaccinations.

Whether Pierce Butler got it **right** really depends on facts we know relatively little about. We don't know in what kind of world we would be living if the western world

had gotten (and stayed) serious about eugenics.

Watchman *says:*

June 4, 2019 at 9:09 am

Or we could just cracking genetic engineering.

I'm also not sure why we need to be concerned about low mutation-selection balance. There's 5 billion of us out there, increasingly connected, so the gene pool is not going to stabilise for a while.

theodidactus *says:*

June 4, 2019 at 10:51 am

Yes exactly. I won't profess to be super biologically literate, but this argument has never made sense to me given the scale of the human gene pool. By the time this would even begin to become a problem, we are quite likely to be either dead or gods. So what's the point?

As for alternate-history eugenics america, I'm not sure what I get from visualizing that, aside from a very interesting alternate history novel someone should write. Eugenics, in the form it took from the 19-teens to the 1950s, just wouldn't work, given our knowledge of genetics today. They pretty clearly thought that "feeble-mindedness" was simple Mendelian trait, like cleft chins, though they got a little loosey goosey when they tried to express it in dominant/recessive terms (citations can be provided if curious). They also usually thought it could be identified in the field by a simple pass/fail test (some of the documents I read trained judges on these tests, which professed to be nearly foolproof and so simple even judges could administer them.)

I don't think a remotely effective "eugenics program" could have emerged before the 1960s or 1970s…if then.

Again, I'm open to someone explaining to the contrary, but as I understand it, we still couldn't effectively select for the kind of characteristics the pro-eugenics progressives thought they could select for, even if we wanted to.

uau *says:*

> *I don't think a remotely effective "eugenics program" could have emerged before the 1960s or 1970s…if then.*

Probably depends on exactly how effective you'd expect it to be, but I don't think it would be that hard to achieve at least some measurable positive effect. There are general things like mutational load that could be improved with many different ways of measuring "badness" even without real understanding of the underlying genetics.

People did successfully breed dogs for various characteristics way before 1960. I don't think breeding humans would be so much fundamentally harder that it'd be impossible before 1960s/1970s understanding.

Watchman *says:*

Dogs were bred for physical characteristics in the main, which were observed to be heritable even without any knowledge of genetics.

Eugenics as discussed here was more concerned with character, and the breeding of dogs has been much less reliable here than in terms of physical appearance. Bad dogs still exist after all. And this is after several thousand years of selection with a new generation every one to

three years (I didn't realise bitches came into best as early as six months!). Human generations are normally in the twenty-year range, so if the analogy to dogs works we're not going to see the desired improvements with any sort of rapidity. We'd also require a socially-dominant idea that not only was eugenics acceptable but that the originally-desired outcomes were still desirable over a span probably measured in the thousands of years; I'd struggle to identify any modern ideology this unchanging over a century, and would suggest ancient ones only appear this way because we can't analysis the fine details of how they operated in practice.

Hide ↑Hide ↑

notpeerreviewed *says:*

June 4, 2019 at 11:49 am

"Eugenics" is a term coined by Francis Galton and he certainly didn't consider ordinary partner choice to be part of it. I'm continually amazed by how broadly people use the term – historically it referred only to intentional efforts to improve the genetics of a population.

Hide ↑Hide ↑

theodidactus *says:*

June 4, 2019 at 11:52 am

Notably though, many progressive eugenicists turned to partner-partner selection counseling after state-backed eugenics efforts failed. It's not a coincidence that Paul Popenoe started the marriage counseling movement.

Hide ↑Hide ↑

notpeerreviewed *says:*

June 4, 2019 at 12:26 pm

And Francis Galton started off by promoting voluntary eugenics, so yeah, voluntary efforts were always part of it. But that's different from genetically-influenced attraction

that's "practiced in every partner choice," which is the misuse I was pushing back on.

DavidFriedman *says:*

June 5, 2019 at 12:20 pm

Galton's essay is [here](#).

He starts off with:

> *EUGENICS is the science which deals with all influences that improve the inborn qualities of a race; also with those that develop them to the utmost advantage.*

All influences include mate choice. Galton goes on to write:

> *The aim of eugenics is to bring as many influences as can be reasonably employed, to cause the useful classes in the community to contribute more than their proportion to the next generation. The course of procedure that lies within the functions of a learned and active society, such as the sociological may become, would be somewhat as follows:*

And point 4 in his list is:"Influences affecting marriage."

So mate choice is part of what Galton included in eugenics.

araybold *says:*

June 6, 2019 at 3:09 pm

Any sufficiently advanced sophistry is indistinguishable from reason.

I don't know enough about that case, but the eugenics movement in general tended to

combine a misunderstanding of genetics and human biology with the assumption that certain ethical positions were obviously correct. In both cases, these views were so widely held that few people realized that the supposedly rational case for eugenics had no solid foundation. This is why one should check one's moral compass before accepting the conclusions of an argument that would be harmful if it turned out to be unsound.
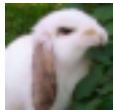
Hide ↑Hide ↑

EmilAich *says:*

June 6, 2019 at 4:47 pm

It's absolutely his religion. Catholicism is strongly pro-life and that means anti-euthanasia, anti-abortion, anti-sterilisation.

https://en.wikipedia.org/wiki/Pierce_Butler_(justice)

Hide ↑Hide ↑

Dan L *says:*

June 3, 2019 at 5:38 pm

> *(This is the correct Bayesian action: if I know that a false argument sounds just as convincing as a true argument, argument convincingness provides no evidence either way. I should ignore it and stick with my prior.)*

I am reminded of a similar sentiment expressed in Hardball Questions:

> *The outside view tells you no; judging from the superintelligence's past successes, it could have convinced you equally well of the opposite position. If you are smart, you will precommit to never changing your mind at all based on anything the superintelligence says. You will just shut it out of the community of entities capable of persuading you through argument.*

It is difficult to articulate how strongly I reject this reasoning.

No, I don't think the false arguments sounds "just as convincing", and I don't think the superintelligence could have "convinced [me] equally well of the opposite position". Maybe my confidence takes a massive hit and I (rightfully!) approach the topic with more caution, uncertain in my beliefs.

But there's an asymmetry there, and [while the storm is strong my shelter *isn't* nothing.](#) And sometimes, [there's work to do](#).
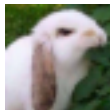
Scott Alexander *says:*

June 3, 2019 at 6:10 pm

Interested in hearing what you think of the next few posts in this sequence!

Dan L *says:*

June 3, 2019 at 7:11 pm

> *this sequence*

!

You had my interest. Now you have my attention.

MawBTS *says:*

June 3, 2019 at 6:12 pm

Given that no superintelligence exists, how can you be sure that one wouldn't convince you?

Bugmaster *says:*

June 3, 2019 at 6:35 pm

Isn't that like asking, "yes but can you *prove* there's no God" ? Explain to me how a superintelligence could even exist in the first place, and how it would convince me of things, and then I'd be willing to consider your argument.

MawBTS *says:*

June 3, 2019 at 6:49 pm

We're talking about whether a superintelligence could convince us of wrong ideas, not whether one could exist.

Bugmaster *says:*

Well, if an omnipotent entity could exist, then obviously it could do anything it wanted, by definition — couldn't it ?

Murphy *says:*

@Bugmaster

Imagine you have a 5 year old child.

You say they should never get in a strangers car.

They argue that they should only get in a car with a stranger who makes a really good case for why they should do so.

How much do you trust a childs ability to reject bad arguments vs an adult trying to be convincing. Note that the child yesterday spent 20 minutes searching for their nose after their uncle "stole" it.

Would they be better off with a set rule of just "never get in a strangers car"? perhaps with an exception for police or something.

Because adults can be really really convincing to 5 year olds.

Now imagine you meet a charismatic genius, as far above you as you are above a 5 year old with a plausible

sounding get rich quick scheme but it'll need your life savings invested...

Bugmaster *says:*

June 4, 2019 at 4:23 am

@Murphy:
Again, I agree that a hypothetical superintelligent entity with the ability to convince anyone of anything would be able to convince *me* of anything. However, you are just positing the existence of such a thing, based on your assumption that the ability to convince people of things can be scaled infinitely high. That's like saying, "I can walk to the store from my house; therefore I could walk across the USA; therefore I can walk to the Moon". It's not enough to just assume your conclusion; you need to provide at least some evidence.

Murphy *says:*

June 4, 2019 at 4:52 am

It may not scale linearly but I suspect we have different intuitions about the difference in direction.

Your intuition seems to be that the difference needs to be massive.

I honestly don't think a real super-intelligence is required. A decent sized group of merely smart people seem to be enough to convince lots of very bright people of fairly absurd things.
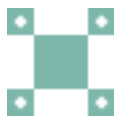
Full grown adults with mostly normal adult cognitive capacity regularly get pulled into cults and fairly obvious scams by someone pressing the right levers.

We don't need to hypothesize a superintelligece when we can just look at regular human level intelligence that are already really really good at convincing people of absurd things.

So while you seem to assume it's home->shops->coast to coast->moon I'm seeing it more like home->shops->shops at the far end of the street->shops a couple of streets over.

Taking the small step of "ok, what happens when you have something even smarter and more capable than the best human conman/cult leader/PR expert" isn't a big leap.

Hide ↑Hide ↑

Jiro *says:*

June 4, 2019 at 8:29 am

> *Full grown adults with mostly normal adult cognitive capacity regularly get pulled into cults and fairly obvious scams by someone pressing the right levers.*

This demonstrates that a certain percentage of adults can be convinced of anything, not that any adult can be convinced of anything. It doesn't follow that all adults are equally vulnerable and it is quite likely that it takes a certain type of person to fall for those.

Hide ↑Hide ↑

deciusbrutus *says:*

June 3, 2019 at 7:57 pm

They could construct an explanation that seemed simple to you that adequately predicts all of your observations to within your ability to measure them, and then extrapolate that model to situations that you never tested where it's wrong.

Except for the superintelligence, that's why people believed that heavier objects fell faster than lighter ones until Galileo. It's why EVERYONE believed that a given power accelerating a given mass yielded an acceleration independent of speed, until Einstein. It's why phlotgetstein theory existed, and it's why there's an example of all of those which I can't name because it hasn't been disproven yet and everyone including me believes it now.

There certainly ARE false things that a superintelligence can convince me of, using methods that appear identical to how a superintelligence would convince me of true things. Therefore, it only way to update correctly is to ask "What fraction of things that this intelligence wants to convince me of are true?" and update towards that.

Hide ↑Hide ↑

localdeity *says:*

June 4, 2019 at 12:18 am

> *They could construct an explanation that seemed simple to you that adequately predicts all of your observations to within your ability to measure them, and then extrapolate that model to situations that you never tested where it's wrong.*

This is not necessarily *possible*. If there's a sequence of ten numbers, and I know the first five and that they fit an arithmetic sequence, it is not possible to construct an equally simple polynomial (i.e. of degree 1 or less) that agrees with those five values but makes a different prediction for the next five. Or—suppose you've thoroughly observed Newtonian mechanics and gravitation in prosaic Earth conditions and have little knowledge of electromagnetism or light, and the AI is trying to convince you of special and general relativity: the Einsteinian theory is *more complex* than the Newtonian when you're only

explaining mechanics and gravitation at nonrelativistic speeds. The Einsteinian theory only becomes less complex when certain empirical observations are made, which you haven't made. The AI would have to either persuade you to take its word on those observations, or persuade you to make those observations yourself. Someone telling you to trust them about a whole bunch of observations is kinda suspicious. Someone telling you to do experiments is less suspicious, although it is possible to mislead someone about how to interpret experimental results (e.g. I've heard of tech support phone scams where the scammer says "Your PC is infected. Open your system log and see all the errors there; that proves it's infected, and the fact I know they're there proves I know what's wrong and can fix it").

You can't necessarily reduce all important theories to a few equations and then count the distinct symbols in the equations. But you could be epistemically cautious with things that *aren't* like that.

Hide ↑Hide ↑

Bugmaster *says:*

I think you might be conflating two positions:

1). A superintelligence could convince me that some false things are actually true.
2). A superintelligence could convince virtually anyone of virtually anything.

Position (1) is very likely, but not very interesting; perfectly ordinary human liars can do that. Position (2) is an extraordinary claim. If I'm standing outside on a bright sunny day, the superintelligence could try to convince me that the sky is pink with white polka dots instead of blue, but it would definitely be an uphill battle — one that I

but it would definitely be an uphill battle — one that I doubt the AI could win just by talking to me.

nadbor *says:*

June 4, 2019 at 6:25 am

FWIW I thought this entire thread was about 1) not 2) I think bringing superintelligence into this was an unnecessary rhetorical flourish. You could replace superintelligence with Velikovsky and the quoted passage would still make sense. Velikovsky won't convince you that 2 + 2 = 5 but if he has a history of making convincing (to you!) arguments about ancient history then it makes perfect sense to commit to not updating too much on the evidence he presents in this domain.

Someone smarter than Velikovsky may have a similar property in a broader domain. An intriguing limiting case is the perfect arguer who can convince anyone of anything but this is really not necessary.

A1987dM *says:*

June 4, 2019 at 6:35 am

[nitpick]You mean force not power — in Newtonian mechanics a constant power results in acceleration inversely proportional to speed[/nitpick]

deciusbrutus *says:*

June 4, 2019 at 4:28 pm

> *This is not necessarily possible. If there's a sequence of ten numbers, and I know the first five and that they fit an arithmetic sequence, it is not possible to construct an equally simple polynomial (i.e. of degree 1 or less) that agrees*

with those five values but makes a different prediction for the next five.

So there is literally nothing that can convince you that a sequence that starts out like an arithmetic sequence might not be one, except direct observation of the next number?

Hide ↑Hide ↑

deciusbrutus *says:*

> *[nitpick]You mean force not power — in Newtonian mechanics a constant power results in acceleration inversely proportional to speed[/nitpick]*

No, I meant constant change in KE, not constant change in speed. Because different non-accelerating Newtonian observers agree about delta Ke and power, but disagree about force.

Hide ↑Hide ↑

localdeity *says:*

> *So there is literally nothing that can convince you that a sequence that starts out like an arithmetic sequence might not be one, except direct observation of the next number?*

Anything that agrees it's an arithmetic sequence for five terms, but disagrees for any later terms, must be more complicated than "P(n) = An + B". It could be *somewhat* simple; but it would be strictly less simple than what I already believed.

Incidentally, my dad had a relatively accessible example for something that exhibits linear response until it doesn't

(after which it follows a different line): shining a light through a piece of paper and measuring how much light goes through. After a certain point, it burns a hole in the paper, and suddenly a lot more goes through. The formula might be a piecewise function; exactly what happens during and after the discontinuity depends on the exact parameters of the experiment (e.g. do you average the amount of light measured over a period of N seconds, or just read the value N seconds after turning on the laser?).

If I didn't know that lasers could burn things, and the superintelligence was trying to convince me of that model… well, it could get me to the point of "Hmm, that sounds plausible. I will want further verification before I say I believe it." (This is my stance on a great many issues.) Does that count as "convincing me"?

Hide ↑Hide ↑
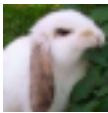
Dan L *says:*

June 3, 2019 at 7:10 pm

Never said it couldn't. The question is whether or not it would find it *equally* easy to convince me in an arbitrary direction. Perhaps the more intelligent than me it becomes the easier it finds the task, but the delta reaches zero only at infinite knowledge.

Or to flip the direction of the problem – that my existing knowledge and rationality has *zero* influence on what ideas in ideaspace I find convincing. To assert that such is the case is to take an absolutist stance on the impossibility of reason. Plenty of reason to reject that.

(Ok, sure – I'm a fallible human and I'm not always going to biased towards the truth. But allowing for meta-reasoning, this is a fixable problem.)
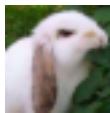
Hide ↑Hide ↑

Murphy *says:*

June 4, 2019 at 4:00 am

non-zero can still be really small.

There's non-zero difficulty involved in convincing a 4 year old that this is a mashmallow farm:

https://i.redd.it/bu9dpgxfdnx21.jpg

but adults will still do so casually, smirking over the kids head because there's a level of effort that's close to free.

Hide ↑Hide ↑

Dan L *says:*

June 4, 2019 at 8:41 am

> *non-zero can still be really small.*

The two differ by a factor of literally infinite, and lead to correspondingly different conclusions. "Really small" bootstraps to "arbitrarily large".

How old is the child when the adult finds the lie impossible?

Hide ↑Hide ↑

DaveK *says:*

June 4, 2019 at 7:47 pm

Whatever, that's clearly a marshmallow farm. Also, where is it?

Hide ↑Hide ↑

Trevor Adcock *says:*

June 4, 2019 at 12:00 am

I could ask it to show me its work. The mathematical proof/computer program that lead it to have that conclusion. If I don't have computational resources equivalent to it, I guess I'm screwed, but that seems more a problem along the lines of a bully being physically stronger than you, than an

info hazard. It could do a lot worse to me if it had those powers than convince me of something wrong.

Randy M *says:*

June 4, 2019 at 7:14 am

People around here have high verbal intelligence, so they have both seen the effectiveness of persuasion, but also to overestimate its scope. I'm not convinced there exists the possibility for a super AI to convince anyone of anything. However, I know there exists a level of intelligence that can convince me of wrong things; indeed, I myself posses such a fearsome intellect.

HeelBearCub *says:*

June 4, 2019 at 7:46 am

> *I'm not convinced there exists the possibility for a super AI to convince anyone of anything.*

I am assuming this just collapses to you not being convinced "AI" can actually exist. Otherwise I have a hard time making sense of it unless it is intentionally hyperbolic?

Randy M *says:*

June 4, 2019 at 8:08 am

Hmm, not sure if you object to what I mean or the equally valid unintended alternate reading that I only see now. "Any" can be ready as "every" when talking about capability, or it can be read as "some, single" thing, which is very nearly opposite.

I doubt the potential existence of an ai that has omnipotent rhetorical powers. I fully expect an ai that can convince many people of many things, some even falsely. Does that make more sense?

HeelBearCub *says:*

June 4, 2019 at 8:18 am

Yes, that clears up the ambiguity.

I had been mistakenly reading you as saying that you thought it more likely that an AI would always be unconvincing.

Hide ↑Hide ↑

Jiro *says:*

June 4, 2019 at 8:31 am

> *However, I know there exists a level of intelligence that can convince me of wrong things; indeed, I myself posses such a fearsome intellect.*

You possess an intellect that can convince yourself of *some* wrong things, not of an arbitrary wrong thing.

Hide ↑Hide ↑

Randy M *says:*

June 4, 2019 at 8:37 am

Right.
In other words, I appreciate Scott's approach despite not believing in unlimited rhetorical powers because I know the limited version is still quite capable.

Hide ↑Hide ↑

Walter *says:*

June 5, 2019 at 8:11 am

"If you are smart, you will precommit to never changing your mind at all based on anything the superintelligence says."

This is the best sentence. I think it is the 'if you are smart' that does it.

JPNunez *says:*

If there is a superintelligence capable of convincing you of anything, who is to say it is going to try arguing at you to do the convincing?

Hell, human level advertising has shifted from writing long paragraphs about the excellences of the advertised products, to blatant putting the product next to attractive people, putting products in movies and tv shows, buying surface of the clothing of famous sports people, etc.

If the paperclip maximizer doesn't have superscience to tile the earth into paperclips, it may try to buy some google ads, instead of trying to reason you into making more paperclips.

Doctor Locketopus *says:*

Someone (I want to say Carl Sagan, but could've been Asimov or possibly even someone else) once noted that when he talked about Velikovsky to historians or Biblical scholars, their take was generally something like "His history is total garbage, but the science is plausible." Conversely, physicists tended to say things like "His physics is absurd, obviously, but his history is intriguing."
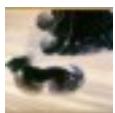
teageegeepea *says:*

That sounds like Clifford Stoll's description of himself in "The Cuckoo's Egg": he lost his job in astronomy and all his colleagues said he wasn't cut out for that, but was a whiz at computers. He then got a job as a systems admin at a university, where his colleagues thought he was ignorant of computers but knew everything about astronomy. This is of course related to the Murray Gell-Man amnesia effect.

Don P. *says:*

June 4, 2019 at 4:43 pm

I think Asimov, based on the rock-solid evidence that I feel that I heard this long enough ago that I remember it from the part of my childhood when I was reading the various Asimov collections of short science-fact columns from...the Magazine of Fantasy and SF? Also, I think Sagan's own background wouldn't require him to check with (other) physicists.

Hide ↑Hide ↑

Concavenator *says:*

June 5, 2019 at 3:37 am

Sagan wrote something like that in one of his books, the physicist being him:

> *I can remember vividly discussing* Worlds in Collision *with a distinguished professor of Semitics at a leading university. He said something like "The Assyriology, Egyptology, Biblical scholarship and all of that Talmudic and Midrashic* pilpul *is, of course, nonsense; but I was impressed by the astronomy." I had rather the opposite view.*

(Carl Sagan, *Broca's Brain*, 1979, Google Books)

(for some reason, I'd have sworn hearing him say this quote in *Cosmos*, but I can't find it in the relevant episode)

Hide ↑Hide ↑
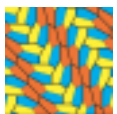
RavenclawPrefect *says:*

June 3, 2019 at 6:10 pm

> *I don't think I'm overselling myself too much to expect that I could argue circles around the average uneducated person. Like I mean that on most topics, I could demolish their position and make them look like an idiot. Reduce them to some form of "Look, everything you say fits together and I can't explain why you're wrong, I just know you are!" Or, more plausibly, "Shut up I don't want to talk about this!"*

I'd love to see some transcripts of in-depth interviews in this situation; what happens when someone without much in the way of epistemic training is posed with devastating counterarguments to their positions? (Not aggressively or anything, just having a conversation

in which one person keeps the discussion pointed towards various pieces of evidence.) How do reactions vary? In particular I'd be interested to see what happens in the case of people whose default is to keep talking calmly rather than to storm out. I assume something like this exists somewhere on the internet (there isn't a shortage of people with silly beliefs or other people willing to debate them), but I've never come across it as far as I can recall outside of televised debates, and those aren't really fair because neither side can possibly concede any ground.

Also, what fraction of people actually end up convinced? I'd guess a negligible amount, but would be pleasantly surprised if this wasn't the case.

HideHide

losethedebate *says:*

June 3, 2019 at 9:27 pm

Jonathan Haidt has done this study. He describes it in his book *The Righteous Mind*, and here is a description online that's more in-depth than the one in the book (full disclosure, I haven't read the link). Basically, he coined the term "moral dumbfounding" to refer to the phenomenon he observed where people would have a moral judgement of a situation, then the experimenter would refute all the person's arguments to the person's own satisfaction, but the person would still keep their judgement, saying things like "I can't explain why, but…".

Hide ↑Hide ↑

Conrad Honcho *says:*

June 4, 2019 at 7:46 am

The historical record shows it works out something like this.

Hide ↑Hide ↑

DragonMilk *says:*

June 4, 2019 at 9:34 am

I'd wager he'd be too polite to say anything at all once he senses defensiveness on the part of the person he's talking to.

Hide ↑Hide ↑

realitychemist *says:*

June 4, 2019 at 1:54 pm

> *Also, what fraction of people actually end up convinced? I'd guess a negligible amount, but would be pleasantly surprised if this wasn't the case.*

This seems like it would be almost impossible to measure. In my experience of having my mind changed I am rarely (although not never) convinced to change my mind on the spot, but a good argument plants a seed of doubt.

We don't usually get to see these seeds grow (or die, as may be the case). We have an argument with someone, plant a seed of doubt, and go our separate ways. They have more arguments, observe the world, have experiences, and change as a person. Eventually they may come to accept that idea you planted (or not), but it's a slow process and it's unlikely you'll ever know unless you do a longitudinal study or something.

But then again maybe this is just me. I don't really know what having their minds changed feels like to other people.

Hide ↑Hide ↑

DavidFriedman *says:*

June 5, 2019 at 12:23 pm

> *I am rarely (although not never) convinced to change my mind on the spot, but a good argument plants a seed of doubt.*

I remember my father saying that the point of an argument is not to convince the other person but to give him the ideas with which he may later convince himself.

Hide ↑Hide ↑

DaveK *says:*

June 4, 2019 at 7:52 pm

This isn't as easy as you think. A lot of subcultures have picked up their own methods of tautological arguments. You're assuming here the person will-1. Argue in good faith 2. Not constantly interrupt 3. Not engage in an endless string of fallaciosu reasoning you have to catch up with 4. Not say things like "this is a conversation, you have to give my point of view equal time 5. Not play "god of the gaps" 6. Not Gish Gallop with

"i'm just asking questions" that never lets you finish a single line of argument.

It's not that these people are genius arguers, its that they learn how to perform this stuff through mimetic repetition. Try "arguing circles" around a committed new ager or SJW type and see how effective that is.

Hide ↑Hide ↑
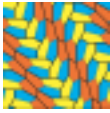
RavenclawPrefect *says:*

June 4, 2019 at 9:03 pm

I agree there are many people for whom this would be totally nonconstructive due to reactions like the ones you mentioned, but I think such people are at least not a large majority of the population of people who could be convincingly argued against. I'm interested in what happens with everyone else.

Hide ↑Hide ↑

Mark Atwood *says:*

June 5, 2019 at 7:47 am

I think such people are at least not a large majority

Such people are effectively ALL the population.

Everyone who isn't is lost in measurement and rounding errors.

Even those very few still are like that most of the time.

Civilization advances at all because such people occasionally find forums together and sometimes haltingly awkwardly error-prone-ly have actual real conversations.

Hide ↑Hide ↑

Ol' Bab *says:*

June 5, 2019 at 11:35 am

I frequent a blog (heavily moderated!) where this comes up occasionally. Various commenters report that if the recipient of the great argument starts to find himself

being convinced, cognitive dissonance/agitation sets in and the discussion is ended. A *very* few are/allow themselves to be, moved in the uncomfortable direction.

The blog is http://www.ecosophia.net, and the commentariat is largely composed of "*people whose default is to keep talking calmly rather than storm out*". The regulars are perhaps 60/40 red/blue, 50/50 pro/con abortion, 80% pro serious climate change. Blogs are weekly, comments in the hundreds (sorry). Main themes are the ongoing, soon to get worse, collapse (driven by resource depletion, the falling of our version of civilization, and dramatic climate change), periodic really wonkish expositions in occult ideas -I skip these-, and really great looks at the craziness of our current political scene.

I apologize for the rather blatant ad for someone else's blog. I of course want to get everyone to join me down the rabbit hole!

Hide ↑Hide ↑

PDF *says:*

June 3, 2019 at 6:11 pm

Hi Scott,

If it's really the case that

> *a false argument sounds just as convincing as a true argument*

then why not change *that*, instead of changing what you do with convincing-sounding arguments? I mean, if rationality is good for *anything*, it's exposing the differences between true and false arguments/claims, right?

I also don't understand how someone as skilled in this as yourself could be considered hopelessly gullible in just one (or a few) arbitrary topic(s). It's been several years since the original post – do you still feel this way? What is it about the topic of pseudohistory that could make one unable to spot false arguments? This sounds a little bit like magic. Have you tried consulting colleagues with the same "insurmountable evidence" and examining their responses?

HideHide

Enkidum *says:*

June 3, 2019 at 6:51 pm

The problem is that, whatever argument-fu you practice, you have a finite supply of

time and there is an infinite supply of bullshit out there.

Hide ↑Hide ↑

deciusbrutus *says:*

June 3, 2019 at 7:59 pm

Inexhaustible, not infinite.

Some infinities can be exhausted, with clever enough algorithms.

Hide ↑Hide ↑

Enkidum *says:*

June 3, 2019 at 8:38 pm

Fair.

Hide ↑Hide ↑

notpeerreviewed *says:*

June 3, 2019 at 9:16 pm

> I also don't understand how someone as skilled in this as yourself could be considered hopelessly gullible in just one (or a few) arbitrary topic(s).

I think I'm hopelessly gullible on *most* topics. Or rather, the only reason I'm not gullible is because I know better than the listen to convincing arguments on topics I know nothing about. If I tried to reach reasoned conclusions about those topics without first spending enough time studying Near Eastern history or whatever, I *would* be hopelessly gullible.

Hide ↑Hide ↑

Clutzy *says:*

June 3, 2019 at 9:19 pm

History seems (to me) to be one of the places your average smart person should be able to dismiss, at least, the worst arguments. Because one need merely to have an accurate model of human behavior to understand that some things are not plausible.

Sometimes, though, there are two histories that are both plausible. Many people thing

Jared Diamond's GG&S narrative of history makes sense. I don't. I do however recognize that most of it has plausible theories that don't diverge from what normal humans would do (my biggest gripes being with his domestication theory and his North American population calculations, the latter of which I think is not consistent with human behavior).

Hide ↑Hide ↑

Watchman *says:*

June 4, 2019 at 12:25 am

You are aware of how extreme the variety of human behaviours have been over time and distance? We tend to view history through the prism of our own understanding. Thus, even if we avoid the trap of moral judgement, we can still fail to actually appreciate how different the thinking involved in something even as recent as say colonialism actually was. Sure we can read the writings of Rhodes or Kipling, or anything on manifest destiny, but we'll not be engaging with them in the same way as their contemporaries, because we don't and can't think like they do: we are post-colonial thinkers due to the age in which we grew up. If we can't fully understand the context of the writers of 100 years ago, how do we appreciate the logic of a mound builder, of a ship burial or of a hermit living up a pole? We can try, and we can construct a comparative picture of their behaviours, but ultimately the major decisions about how we understand any historical behaviour are (currently) made in twentieth-century minds. Your construction of an accurate model would, however hard you tried, be an attempt to filter history according to what seems plausible to you in the here and now.

Hide ↑Hide ↑
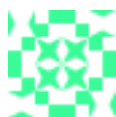
Clutzy *says:*

June 4, 2019 at 1:23 pm

I'd say that your theory of a complex evolution of the state of humans is the kind of naivety that I presume causes historical illiteracy.

I have no trouble understanding the actions of colonials, or slavers, or Alexander the Great. They all code to me as humans doing human things when presented with human circumstances.

Things that code as false to me is when a historian presents "kumbayaa" scenario.

Hide ↑Hide ↑

Matt M *says:*

June 4, 2019 at 7:55 am

> *Because one need merely to have an accurate model of human behavior*

This seems easier said than done.

I know a whole lot of really smart people who would disagree *strongly* in terms of how they "model" human behavior...

Hide ↑Hide ↑

Clutzy *says:*

June 4, 2019 at 1:24 pm

True, but that is not merely present in those people's evaluation of history, it also is shown in their evaluation of the present and future.

Hide ↑Hide ↑

MawBTS *says:*

June 3, 2019 at 6:23 pm

> *then why not change that, instead of changing what you do with convincing-sounding arguments? I mean, if rationality is good for anything, it's exposing the differences between true and false arguments/claims, right?*

Because it's impossible: nobody can be perfectly informed on every single topic under creation. There's too much to know.

Even if Scott spent four years getting a doctoral-level education in history, he'd still be gullible on household plumbing (for example).

HideHide

DavidFriedman *says:*

June 3, 2019 at 8:11 pm

You don't have to be perfectly informed.

Scott's claim is not merely that false arguments are sometimes as convincing as true ones but that the probability of an argument being convincing conditional on its being false is the same as conditional on its being true. I think that's clearly wrong.

Hide ↑Hide ↑

Athenae Galea *says:*

June 4, 2019 at 2:40 am

While I'm not sure I agree with the strong form of what Scott has said either, I would definitely not go as far as "clearly wrong". While the claim as taken literally *is*, I think a more reasonable interpretation would be "if I know that a false argument**, conditional on my having heard it from a convincing source,** sounds just as convincing as a true argument, argument convincingness provides no evidence either way"
The claim isn't (I'm imagining) that the average convincingness of all imaginable wrong arguments is the same as of all imaginable correct arguments, because that, as you say, would be obviously wrong.

Some of the examples given are ones where *both sides* are able to be absolutely convincing, and so in that family of cases anyway the claim is trivially true. This is effectively "I am currently convinced, however I am aware that if I read the rebuttals to these my certainty would drop off a cliff, so I might as well throw it off now and save myself the trouble." So to speak.

Hide ↑Hide ↑

BlindKungFuMaster *says:*

June 4, 2019 at 5:18 am

Liars (or motivated reasoners) have a lot more degrees of freedom.

Hide ↑Hide ↑

baconbits9 *says:*

I don't think that your position holds. Scott's position is more like 'a false argument that has been adopted and is being presented to you is as likely, or close to as likely, to be convincing as a true one', and if you were to flesh it out you would get something like 'given that there is one correct position and a near infinity of false ones the most convincing false arguments are going to sound as convincing as the true ones.'

Hide ↑Hide ↑

emiliobumachar *says:*

June 5, 2019 at 2:43 am

Consider survival bias. The probability of a *surviving* argument being convincing conditional on its being false is the same as conditional on its being true. Not that clearly wrong, is it?

Hide ↑Hide ↑

Scott *says:*

June 3, 2019 at 6:28 pm

Two posts that I think pair really well with this one, looking at the failure modes of taking ideas too seriously:

* Reason as Memetic Immune Disorder, by Phil Goetz
* Nerds are Nuts, by Razib Khan

HideHide
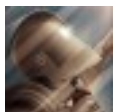
JohnBuridan *says:*

June 3, 2019 at 6:32 pm

Still good. Glad you reposted.

A student asked me last year: how does one prove that say, the American Civil War, happened to someone who denies it? I said you both need to agree to standards of evidence which will convince both of you one way or another. But I went on to point out that frequently people have other reasons for wild beliefs that are not about evidence, but some other issue.

But By Abraham! How delicate is the balancing act!

I have been interested for some time in creating a small course on coping with epistemic

uncertainty, conspiracy theories, and standards of evidence.

AZpie *says:*

June 3, 2019 at 8:16 pm

Thought of this a lot recently, and ended up going in circles. A very, very big deal in systems such as standards of evidence and such is the fact that nobody can draw or redraw the map for science alone. It's ultimately a matter of trust, and epistemically I find that to be a big problem. There's no way I can ever analyze all the evidence by myself, and if there exist systematic cognitive, social or ideological biases affecting the production of information, the way it's discussed and whether it's going to be ultimately prioritized or dismissed, I'll have no hope at all of ever discerning legit information from potential bullspit.

That is, everything is potential bullspit, which is a wonderful doctrine – *constant vigilance! Always doubt*! – except that it also leads to historical revisionism, climate change skepticism, anti vaxxxing and nice stuff like that. So I'm back to square one: there must exist a system of thought with which I can deal with the information at hand and grade beliefs according to the evidence supporting and rejecting them. But if I really think it through...

I do think that that's the best we've got at the moment. But if I Take The Idea That I'm Gullible And Not An Expert Seriously, I'm going to run into the weeds – very, very deep.
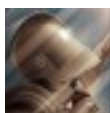
JohnBuridan *says:*

June 4, 2019 at 6:57 am

It's such a vicious and difficult position that one realizes how horribly dependent even that smartest people are. David Foster Wallace's introduction to Best American Essays 2007 is a fantastic little meditation on this problem. http://neugierig.org/content/dfw/bestamerican.pdf

Quotes:

> *The general point is that professional filtering/winnowing is a*
> *type of service that we citizens and consumers now depend on*

> *more and more, and in ever-increasing ways, as the quantity of available information and products and art and opinions and choices and all the complications and ramifications thereof expands at roughly the rate of Moore's Law.*
>
> *To really try to be informed and literate today is to feel stupid nearly all the time, and to need help.*

Hide ↑Hide ↑

g *says:*

June 5, 2019 at 7:04 am

> *How delicate is the balancing act!*

Username [checks out](#).

Hide ↑Hide ↑

Bugmaster *says:*

June 3, 2019 at 6:39 pm

Maybe it's because I'm a contrarian by nature, but still, I've never found this argument particularly insightful. If you discover that you can be persuaded to believe in mutually contradictory positions, then maybe you need to examine the flaws in your reasoning that are preventing you from forming correct conclusions — as opposed to just shutting out anyone who attempts to convince you of anything. This is especially important if one side is trying to convince you of A, and the other one of not-A, because in this (admittedly, rare) case, one of them has got to be right.

HideHide

deciusbrutus *says:*

June 3, 2019 at 8:02 pm

I've seen lots of cases where one side is lying to convince me of A and another side is lying to convince me of not-A and another side is lying to convince me to reject the dilemma.

And in fact, the correct answer is "A and A is irrelevant to the thing that all of you are

applying it to . And also "Screw all of you who are lying to convince me of something."

Bugmaster *says:*

June 3, 2019 at 8:32 pm

Well yes, but this is not Scott's scenario. In his scenario, he explicitly cannot identify which (if either) side is lying; at least, not right off the bat.

deciusbrutus *says:*

June 4, 2019 at 4:46 pm

It is not the case that one or more of them has to be right.

AZpie *says:*

June 3, 2019 at 8:25 pm

I don't think the issue is merely flaws in one's reasoning, but instead the limits in one's abilities to effectively process information. "Let's assume a perfectly rational, infinitely intelligent being…" With the Art, we can actually try to approach the first, but the second part is what really gives us problems. In most complex issues (say, nearly all political economics) I've found so many convincing arguments for both (all) sides I've lost count. I can't keep up with it unless I devote all my life to a single question alone – and it's not the only question that bothers me. The issue is not that my reasoning is wrong, but that I lack knowledge to relate the arguments to one another to compare and weigh them relative to the whole of facts, evidence and discussion. I can compensate to an extent by being systematic and thorough, but it won't get me indefinitely far.

Bugmaster *says:*

June 3, 2019 at 8:36 pm

I don't think you need to be (or even posit) an infinitely intelligent being in order to reason in a rational fashion; in fact, we wouldn't need the Bayes rule if we were already omniscient. Nor do you need perfect knowledge in order to form a practically applicable opinion.

For example, when I'm buying a car, I'm perfectly fine with being 80% or so convinced that the car I chose is the best one to buy. Yes, I could be wrong; if I studied automotive engineering for 20 years then obviously I could be better informed — but the stakes are low enough so I don't need to do that. If I narrowed down the choices, and it came down to a 50/50 split between two cars, I would either postpone my purchase so I could learn more, or flip a coin. I wouldn't just go, "oh well, no car for me".

Hide ↑Hide ↑

AZpie *says:*

June 4, 2019 at 6:54 am

Thanks for replying.

One third of my "Let's assume…" was a joke. A third of it was meant to point out that a confident epistemic status requires not only a careful analysis of evidence available to one's mind (the Art), but also faculties capable of processing that information (smarts and knowledge). A third of it was to poke at the fact that while attempting to be rational about things and use maths to improve our decision making, we're bound to make assumptions, not all of which are immediately obvious and not all of which are, you know, smart. My point was not to imply that rational or reasonable decision making would take infinite brain power.

Practical applicability is also a matter somewhat distinct from epistemic uncertainty, but relevant enough here. However, I'd like to point out that the fact that we need to operate somehow with uncertainty does not mean that we are making good decisions, and the results of our decisions do not necessarily tell us whether we could have done better or worse – moreover, the result of our actions does not necessarily relate to our knowledge in any way. (say, I'll decide to invest to a stock because God told me it'll get a pump. It does. Did I know that'll happen? This isn't a very applicable example for the discussion at hand, but I just wanted to point out that epistemic (un)certainty and (apparent) practical

applicability need not go hand in hand.)

As for your example, I agree with you. Buying a car that suits my needs well enough is, however, a relatively simple situation, especially considering the fact that whether or not the car I end up buying actually is "the best for my needs" or whatever will most likely never become apparent to me, unless it suffers from some serious flaws or I'll have some other means of finding out eventually. If I don't, I might just assume that since I'm content with the car, I probably got the best one. I won't *know*, though.

However, most complicated questions aren't such that I could assume an 80% probability of ending up with the best solution. They are such that most the times I can't give any probability estimate for being right, unless I just pull one out of my arse to look cool.

For example, I might consider which strategies Finland should choose to reduce unemployment and improve the well-being of relatively poor people. Mostly people just decide to side with a camp (right-wing, libertarian / left-wing, welfare-statist) and thereby reduce the number of possible solutions. Then they take a look at the proposed political solutions and pick one that sounds good enough to them. We can agree that that's not rational. Being rational in such a matter is a humongous task. The amount of data one should actually consume is absolutely astonishing. We might contend that "for practical applicability, we'll settle with x amount of hours put into it…" but that doesn't tell us anything about our epistemic status. We might make assumptions that the data provided to us is valid, that the statistical analyses others make of the data are valid, that we aren't being lied to, and that we have enough relevant data to form an opinion which won't be shifted so much that we'd care.

All of those are assumptions which do not tell us anything about our epistemic status on the matter. They are a way for us to input "exit program" so that we'll have something nice to do with our time and so that we won't starve while thinking.

No matter how I look at it, it seems to me that it boils down to trust between people; nobody's smart or fast enough to do all the science or Bayes or whatever by themselves, so we need to cooperate and believe in that cooperation. To me, that sounds like I'm never going to really be confident in my beliefs. To someone else it'll sound like something different.

What you're saying about practicality certainly applies, and that the sort of ultimate skepticism I'm presenting doesn't really provide anything worthwhile insofar as we're actually trying to solve problems. However, when it comes to the question of whether or not I know spit, it matters.

Hide ↑Hide ↑

HeelBearCub *says:*

June 4, 2019 at 5:23 am

Despite Scott coming out of the Less Wrong, Bayesian, rationalist community, his intuitions don't seem to actually run that way. He wants certainty. He wants to optimize for all positive outcomes (not sure if optimization is LW thing but it seems like it would be.)

It strikes me that there is a phrase in the rationalist community that is actually very non-Bayesian that actually captures this problem. "If something is true, I want to believe it to be true. If something is false, I want to believe it to be false." Sure, it's a good principle, but at heart it establishes a binary of truth and falsehood as a certainty. The idea of something being, say, "untrue but not false" isn't captured.

Not sure where I am going with this, but it seems like there is some excessive rigidity here that denies necessary flexibility.

Hide ↑Hide ↑

Aapje *says:*

June 4, 2019 at 5:42 am

You can be fairly certain about probabilities. Being certain that a dice has an equal chance to land on each side, for example, which is very relevant if one

decides to wager on the outcome of a dice throw.

HeelBearCub *says:*

June 4, 2019 at 6:00 am

That is a facile and simple example, which isn't really respondent to my prompt. I'm talking about the end results of highly complex systems.

It's also, as stated, something that's untrue, but not false.

Dice games are almost always, hell, always, played with participants that are not "certain" the die sides are all equally likely. You've hid the ambiguity using the word "fairly", then immediately dispensed with it and used "certain" without modifier.

Nonetheless, dice game participants are highly inclined to believe that the dice are fair. ("Fair" being a better word than "equal"). However, it wouldn't surprise me to think that the the more one plays, the less confident one is in the fairness of the dice.

This is due to more frequent players being more likely to displaying addictive behaviors, and the increased probability they have ever encountered loaded dice by virtue of playing so many games.

So, we can put all these modifiers on the initial statement, yet it remains relatively true.

Aapje *says:*

June 4, 2019 at 7:31 am

"Untrue, but not false" seems to be a case of asking the wrong question, which is a topic that Scott has frequently addressed (just like 'true, but not correct').

Crafting the kind of hypotheses that can be answered fairly

(sorry) unambiguously is at the core of science. Anchoring your beliefs on those seems like a good idea to me, even if some of the answers you want can't be answered fairly unambiguously.

Hide ↑Hide ↑

HeelBearCub *says:*

June 4, 2019 at 8:13 am

It can be asking the wrong question. But, again, if we are talking about very complex systems, it might be asking some right question at the wrong time.

Is homeopathy false? But what about inoculation? How about the practice of taking patient histories?
Is leaching quackery?
Is Lamarckian evolution disproven?
Is junk DNA actually non-functional?

But, this is all something of a digression. Because what I was talking about wasn't Scott's ability to reason through these things, but rather his intuitions. There is a mindset that is really uncomfortable with holding these kinds of ambiguities in mind. They want it to be binary, even if at a rational level they understand it is not.

Hide ↑Hide ↑

Aapje *says:*

June 4, 2019 at 12:36 pm

People who are truly uncomfortable with ambiguities are tankies, stormfronters or other extremists. Scott is at most mildly intolerant, which doesn't seem like a bad habit to me.

If your complaint is that Scott isn't perfect, then I'd agree, but this merely makes him human.

sclmlw *says:*

June 4, 2019 at 9:01 am

Agreed. Mathematically, once you have 100% certainty you can no longer do Bayesian statistics on a thing. I feel like there are a lot of Bayesian 'fans', talking about updating their priors based on evidence, but true Bayesian statisticians can't become 100% certain of anything.

That's said, what Scott outlines is a simple case of setting very high probability priors for the standard narrative in any fields you aren't an expert in. This itself is probably loosely based on a kind of fan-Bayesian learning-from-experience that in complex fields you can be convinced about anything by being presented a biased selection of evidence. It's perfectly rational to ignore argument you're not ready to analyze for accuracy, and entirely susceptible to Bayesian analysis. But that doesn't mean most people are applying either standard.

DragonMilk *says:*

June 4, 2019 at 9:37 am

How are you so sure your reasoning is reasonable?

MawBTS *says:*

June 3, 2019 at 6:40 pm

> But I'm also glad epistemic learned helplessness exists. It seems like a pretty useful social safety valve most of the time.

In first aid classes, they give you trial scenarios. A common one is a collapsed stage at a rock concert. 20-30 people are lying around, in uncertain states of health. There could be broken bones, spinal injuries, concussions, and more. As a first responder, how do you ensure that everyone gets the care they need?

You don't. There's no time. While you're doing CPR on one person, six more might be bleeding out. So you just do the best, quickest job you can: make sure each person's airway is clear

check for broken bones, and then put them in the recovery position. You do this for everyone, and only when it is done do you examine individual casualties in more detail.

There's so much information out there that often I adopt a similar, one-size-fits-all response to new theories. "Wow, that sounds interesting. I should learn more about this some day." And then I continue believing what I believed before, unless there's a particularly reason to do otherwise.

HideHide

deciusbrutus *says:*

June 3, 2019 at 8:03 pm

Triage of arguments?

Hide ↑Hide ↑

Clutzy *says:*

June 3, 2019 at 9:21 pm

This is the most brilliant post I've seen this year.

Hide ↑Hide ↑

baconbits9 *says:*

June 4, 2019 at 6:29 am

It's a good approach and it gets easier when you realize that you don't need an opinion on most things. What caused the bronze age collapse is an interesting question with some neat arguments but its not actually important that I pick an explanation as 'correct' outside of some very specific circumstances (ie I'm a history professor).

Hide ↑Hide ↑

Matt M *says:*

June 4, 2019 at 7:58 am

> *What caused the bronze age collapse is an interesting question with some neat arguments but its not actually important that I pick an explanation as 'correct' outside of some very specific circumstances*

This post would have done me good some time before I spent my fifth hour watching YouTube videos regarding the plausibility of "Indoctrination Theory" to explain the ending of Mass Effect...

Spookykou *says:*

I've never understood why someone would want to believe the Indoctrination Theory, head canon should be something better than what happened, but the Indoctrination Theory is even worse than the canon ending at undercutting what I liked about Mass Effect. Mass effect for me was all about making important choices that mattered, and the official ending undercut that in a number of ways that I didn't like, but the Indoctrination Theory calls every 'choice' you ever make into question because you are being indoctrinated, it's throwing out the baby with the bathwater.

Matt M *says:*

It's been a few years, but my understanding is that even within the subset of people who believe in IT, there are strong disagreements as to exactly *when* the "I" really kicks in, up to and including "immediately before the final ending sequence"

In any case, I'd suggest the reason people "want" to believe it is that it strongly implies the writers knew what they were doing and crafted a complex and legitimate narrative all along, rather than threw some hot garbage together at the end in order to make a tight release deadline (also known as the Occam's razor theory of anything about a videogame you don't like)

Spookykou *says:*

June 4, 2019 at 11:52 am

Yes I think you are correct that a lot of people probably like it because it makes Mass Effect a better 'book', I just never thought it was a good book to begin with, it reminds me of fondant on cakes, making food taste worse so it can look better is a ~~failure mode~~ choice that I can only grasp in the abstract.

Hide ↑Hide ↑

Matt M *says:*

June 4, 2019 at 12:03 pm

Also, the question of "Why wouldn't the reapers just indoctrinate Shepard" is almost certainly one of the biggest plot holes in the entire game.

I believe the canon explanation is something like "He/she has a really strong mind and resists it!"… which seems superficially okay, except that it can be presumed that Saren, Benezia, and TIM (YMMV on whether he ended up indoctrinated or not) certainly had really strong minds too, and it didn't do them much good.

Hide ↑Hide ↑

Furslid *says:*

June 3, 2019 at 6:48 pm

Simply recognizing a good argument against a consensus is not enough. It is also necessary to understand the arguments for the consensus. If someone only investigates the arguments for one position in a debate, they aren't following good epistemological practice.

One has to be a good enough historian to understand the arguments for the commonly accepted history before they are competent to evaluate arguments for an alternative history.

The right response to many of these arguments is "That is a very interesting argument, but I do not have enough understanding of the field to evaluate it."

LadyJane *says:*

June 3, 2019 at 7:03 pm

The trick is coming up with a good meta-argument for why your own unwillingness to change your views based on arguments is simply good epistemic hygiene, while everyone's else unwillingness to change their views is foolish irrational stubbornness. That way you can have your cake (refusing to change your views) and eat it too (presenting yourself as superior to others who refuse to change their views and publicly criticizing them for it). It seems like a difficult task to pull off without looking like a hypocrite, but I'm sure you can find some argument to justify it, if you're skilled enough at rhetoric! (Of course, I'm using "you" in the general sense, I don't think that Scott would be particularly inclined to pursue this course of action.)

deciusbrutus *says:*

June 3, 2019 at 8:07 pm

That course of action is required; every culture that does not cultivate resistance to changing their views will find all of their members taken over by someone from a culture that cultivates using the most effective methods available to change other people's mind and also practices resistance to that.

Meanwhile every culture that for moral reasons rejects effective ways of convincing others will find itself crowded out by cultures that do not, regardless of their resistance.

Moloch wins.

LadyJane *says:*

June 4, 2019 at 2:43 am

I agree that it's necessary; as you say, *someone* is bound to do it (and in fact, many people and organizations are already doing it). If it's going to happen regardless, then I have an interest in making sure that it happens for the best possible set of views (as defined by me).

deciusbrutus *says:*

That's what my closing was saying.

PeterDonis *says:*

@Scott:

I think the reason epistemic learned helplessness is a workable strategy for most people is that most beliefs that various people will try to argue you into or out of don't really matter anyway in terms of practical consequences. Suppose it turned out that Velikovsky was right and Venus was originally a comet. How would that affect the practical decisions you make? Not at all, unless you were someone who had invested a lot of effort in arguing that Velikovsky was wrong, and even then the consequences would be in a narrowly limited area of your life–they wouldn't affect how you bought groceries or what you did when your car needed repairs, etc. In other words, you didn't so much accept the arguments of the experts in Near East history, as not care enough to even bother investigating. It's not like you went out and started proselytizing against Velikovsky and calling out people who refused to accept the standard account of experts in Near East history.

So if epistemic learned helplessness is indeed a good strategy for many people (and it certainly has worked for me in the past, although I didn't give it that name), it would seem to imply a higher-order strategy for societies: that it's good to minimize the set of situations where people cannot simply refuse to accept arguments and go on with their lives without having to make a commitment at all on the issue in question. And it also implies a corresponding higher-order strategy for individuals: that you should distrust people who try to argue that you *must* commit yourself one way or the other on whatever their pet issue is–people who say their issue "demands action", etc.

Bugmaster *says:*

While what you say is true, determining which beliefs do or do not matter can get tricky.

For example, does it matter how old the Earth is ? Not really; it could be 6,000 years or 100,000 or 4.5 billion; to the average person all those numbers just translate to "practically infinite". But if you want to have modern science and especially medicine, then you pretty much have to settle on one of those options and ignore the others — and you have to do this as a society, because you want to make sure that people who are capable of pushing the boundaries of science and improving medicine have ample opportunities to do so.

Hide ↑Hide ↑

PeterDonis *says:*

June 3, 2019 at 7:29 pm

if you want to have modern science and especially medicine, then you pretty much have to settle on one of those options and ignore the others

Remember we're not talking about a young Earth creationist (to use the example you give) here. We're talking about someone who is being talked at by a YEC and finds their arguments convincing and is trying to decide whether to believe them and modify their actions accordingly.

If the person is, say, an archaeologist or an evolutionary biologist, then of course it matters which option they choose. But such a person is not going to find the YEC's arguments convincing in the first place. To an average person, it doesn't really matter how old the Earth is as far as their everyday life is concerned. So they are perfectly fine practicing epistemic learned helplessness in this case.

you have to do this as a society, because you want to make sure that people who are capable of pushing the boundaries of science and improving medicine have ample opportunities to do so

That's what my meta-level strategy is for. As a society, we should simply not listen to YEC's who insist that society will collapse and the world will end if their beliefs are not adopted, that it's impossible to be neutral about it, that anyone who does not accept their beliefs is obviously deranged and evil, etc.

In a society that does that, YEC's will be unable to prevent people who want to improve science and medicine from doing so.

Bugmaster *says:*

> *In a society that does that, YEC's will be unable to prevent people who want to improve science and medicine from doing so.*

I disagree, because beliefs are all linked together. If you believe that the Earth is 6000 years old, then you have to either disbelieve in evolution, or do some pretty heavy mental gymnastics. Both of those solutions pretty much preclude you from advancing modern biology and medicine. Obviously, a few people would still be able to do so, but many more would probably go and do something else.

For example, they could study physics… except not really, because of radiometric dating. They could become computer scientists, but computers are built using our knowledge of physics, so that could be an issue. They could Of course, they could become poets or theologians; but at that point, neither medicine nor physics nor a myriad other things are getting done at full capacity.

PeterDonis *says:*

beliefs are all linked together

To the extent this is true, as I said, it means the premise of Scott's argument can't be satisfied–because anyone hearing an argument for, say, young Earth creationism will have so many other beliefs that are related somehow to the age of the Earth that they will easily be able to find a flaw in any YEC argument.

However, I'm not convinced there actually is this much linkage in most people's beliefs. See below.

If you believe that the Earth is 6000 years old, then you have to either disbelieve in evolution

Which indeed most YECs do, as far as I can tell. In fact, YECs are basically applying epistemic learned helplessness in the opposite direction, so to speak: they believe the Earth is 6000 years old because their religion tells them so, and therefore they simply refuse to listen to arguments and evidence for evolution, because they know there must be a flaw in them somewhere but they don't know if they would be able to find the specific flaw in every such argument.

or do some pretty heavy mental gymnastics

Which we have abundant examples of people doing in human history, so this is by no means implausible.

Both of those solutions pretty much preclude you from advancing modern biology and medicine.

Yes, that's true. And as far as I know, there are very few, if any, YECs who have made advances in these areas. But I don't see what this has to do with Scott's argument or my extension of it. There are lots of people who aren't YECs and who are making plenty of advances in science and medicine; they're the ones who might benefit from a strategy of epistemic learned helplessness when YECs try to proselyze them, since it would spare them the mental effort of having to try to spot the flaws in the YEC arguments, which is mental effort they could put to better use advancing science and medicine.

Hide ↑Hide ↑

Bugmaster *says:*

@PeterDonis:

It actually sounds like you mostly agree with me. Still, I'd like to correct one point early on:

> *However, I'm not convinced there actually is this much linkage in most people's beliefs.*

This depends on what you mean by "beliefs". If you really understand how science works, then you currently have very little choice but to believe that radiometric dating (as well as other types of dating) is accurate. There really aren't two separate, mutually independent and/or weakly linked beliefs such as "the Earth is old" and "radiometric dating works"; rather, there are a couple of equations and a bunch of data that leads you to both conclusions simultaneously... plus many others besides.

> *There are lots of people who aren't YECs and who are making plenty of advances in science and medicine; they're the ones who might benefit from a strategy of epistemic learned helplessness...*

My point is that there's nothing magical about this strategy that makes it apply just to the people whose beliefs are already correct. If everyone applied epistemic learned helplessness, then we'd have many more YECs than we do now — especially since YEC beliefs are so much more emotionally satisfying than the alternatives. This means that some people who could've become biologists would become YECs, instead.

Hide ↑Hide ↑

Randy M *says:*

I think you could entirely deny evolution and still contribute in the fields of medicine or computers, or even physics. Knowledge is connected, but usually only by distant implications that don't impact the activities needed to progress research.

At the highest levels, you might be constrained in the correct hypotheses you consider for some phenomenon if you believe the earth is <10,000 years old, but you can still run a drug trial, study metabolic pathways, learn a program a telescope, or whatever.

Understanding evolution is important to biologists, but practically it doesn't impact everything they do. It's more important in satisfying their motivation to understand and give context to their work.

It's not an argument for being wrong, or even not arguing about what is right, but knowledge can be advanced even if some of the people researching don't grasp the whole of it correctly.

Hide ↑Hide ↑

John Schilling *says:*

> *I disagree, because beliefs are all linked together. If you believe that the Earth is 6000 years old, then you have to either disbelieve in evolution, or do some pretty heavy mental gymnastics.*

The mass of the Earth is approximately six billion trillion tons. Every year, the mass of the Earth increases by approximately thirty thousand tons. Therefore the Earth must be approximately two hundred thousand trillion years old. To believe that the Earth is a mere four or five billion years old, you either have to disbelieve in micrometeorites or do some pretty heavy mental gymastics. Fortunately, such "Young Earth Impactionists" will never leave the

Earth, because disbelieving in micrometeorites means that any spaceships they build will be inadequately protected against micrometeorite damage.

Meanwhile, here in reality, actual YEC biologists, doctors, etc, seem to have little difficulty believing in e.g. the development of antibiotic resistance, or most other aspects of evolution relevant to the practice and study of medicine in the post-Noachian era. You are free to call whatever it is they are doing "heavy-duty mental gymastics", but fine, they're all heavy-duty mental gymnasts. Just like almost everybody else, probably including yourself.

eyeballfrog *says:*

> *For example, they could study physics... except not really, because of radiometric dating.*

Not so. There are two easy methods to deal with this.

1) Believe that, for some reason, the world was created in such a way that it looks like it's 4.5GY old. Your model is now consistent with radiometric dating, and any questions about *why* it was created that way can be easily shrugged off.

2) Don't think about it at all. The vast majority of physics does not involve radiological dating. It is perfectly possible to study quantum field theory, atomic physics, or plasma dynamics without ever even measuring a radionuclide.

PeterDonis *says:*

If you really understand how science works, then

If you really understand how science works, then…

Then there are an awful lot of beliefs that, as you say, you're forced to accept whether you like it or not. However, as others have pointed out, people seem quite able to compartmentalize their beliefs so that they can contribute to one area of science without having to face the fact that their beliefs are inconsistent with science as a whole.

But anyway that's not the case I'm talking about. The case I'm talking about is a person who doesn't really understand how science works, but who can get away with that because, as far as their life is concerned, they don't have to understand how science works. They don't want to contribute to science, they just want to make use of its products. Modern technology allows a huge number of people to make use of the products of science while having no idea how they work, and while having beliefs that, though they don't realize it, are inconsistent with the fact that their products work. Their smartphone works even though they believe things that contradict quantum mechanics, and their GPS works even though they believe things that contradict relativity. As long as their beliefs don't lead to obviously counterproductive actions like jumping off a cliff, they have no practical consequences. And so they can simply refuse to accept it when someone tries to convince them that the Earth is 4.6 billion years old, or that light rays recede from you at the same speed no matter how fast you run after them, etc.

Hide ⇑Hide ⇑

Peter Gerdes *says:*

First, (outside of some special counterexamples involving the manipulation of inconsistent belief sets via choice of argument presented) people should believe the conclusions of **actually** valid arguments provided they believe the premises. What we are really interested in is whether people should believe the conclusions of argument which **seems** valid to them.

I think the answer here is very much yes as well. It's just that in evaluating whether an argument seems valid one isn't restricted to looking at the logical structure or the evaluating it step by step. One also considers questions like "Do experts I'm inclined to trust believe this", "If this was compelling would I expect it to be widely accepted or at least recognized as a serious possibility." Many people in areas they aren't super well-informed in or realize they aren't great at very much do exactly this. Indeed, I think most people do this by default anyway so it's not a serious concern.

Where people do fall down is that they don't even appropriately update on the seemingly valid argument after taking account of the countervailing factors like plausibility, expert opinion etc.. and in that failure I very much agree with your friend.

HideHide

tossrock *says:*

June 3, 2019 at 7:37 pm

No one's given a coherent counter-argument to the simulation argument? Really? How about "Each layer of simulation introduces an exponential slowdown which puts an upper limit on the possible depth of the recursive simulation stack when assuming finite time & computational resources"?

HideHide

Doctor Locketopus *says:*

June 3, 2019 at 7:56 pm

There's no (obvious) way of telling how slow (or fast) the simulation is running from *within the simulation*. Even if each picosecond "tick" (as seen by the creatures within the simulation) took a billion years as measured in the world that was *running* the simulation, all would still appear to be normal to the creatures *within* the simulation.

With respect to available resources and time, we have absolutely no basis for judging whether any such limitations apply within the world *running* the simulation. The laws of physics and indeed mathematics itself might be different in such a world.

Recommended reading: Permutation City by Greg Egan.

Hide ↑Hide ↑

Bugmaster *says:*

June 3, 2019 at 8:39 pm

How about the simpler argument ? If we cannot detect whether we are in a simulation or not, by any means, then it doesn't matter and we might as well assume the simpler scenario — that reality is actually as real as it appears to be. On the other hand, if we can detect this, then let's stage some experiments, collect some evidence, and determine whether we live in a simulation or not. Until then… we might as well assume that reality is as real as it appears to be, because it makes the math easier.

Hide ↑Hide ↑

helloo *says:*

June 4, 2019 at 7:38 am

But to a lot of people, it is simpler to assume we live in a simulation than otherwise.

I mean, why would there be a maximum speed? A minimum distance? Things that have probabilistic values until they are observed?
To many people, "we live in a simulation" is an answer to that that is a lot more compelling and simpler than "we don't know" or "that's just how physics is".

I'm not quite sure how the math becomes easier if we don't live in a simulation without some assumptions about what exist outside of it.

Hide ↑Hide ↑

eyeballfrog *says:*

June 4, 2019 at 12:07 pm

> *I mean, why would there be a maximum speed?*

Why *wouldn't* there be? The underlying symmetries of relativity are quite elegant, and there's no real reason the universe wouldn't exhibit them anyways.

> *A minimum distance?*

This isn't known to be true, and is currently inconsistent with general relativity. It definitely makes the math more complicated, not less. **R** is a simpler structure than **Z**.

> *Things that have probabilistic values until they are observed?*

This also isn't known to be true. Bohmian mechanics still isn't ruled out, though it would be if a minimal length scale were discovered. Further, it's very unclear to me how this supports the simulation hypothesis.

Hide ↑Hide ↑
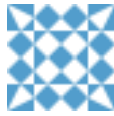
jermo sapiens *says:*

June 4, 2019 at 12:16 pm

> *I mean, why would there be a maximum speed? A minimum distance? Things that have probabilistic values until they are observed? To many people, "we live in a simulation" is an answer to that that is a lot more compelling and simpler than "we don't know" or "that's just how physics is".*

I fail to see how living in a simulation is an answer to why objects cant move faster than the speed of light. In a simulation, it's alot easier to program Newton's equations than Einstein's. Not that I pretend to know what's going on in the minds of our simulator overlords.

But if you find "we live in a simulation" to be a compelling answer to anything, maybe rethink a few of your basic assumptions.

Hide ↑Hide ↑

helloo *says:*

As I mentioned, "that's just how physics is" is a perfectly acceptable answer for why certain things are the way they are.

There is nothing about those things that necessarily point to being in a simulation, but some of those things are similar to features that a simulated reality in our world might include to say deal with the issues of complexity and trying to display an infinite continuous simulation with limitations in the simulation technology (pixels, max int value).

The counterargument is then that a lot of these things make it a lot less elegant and add additional complexity to what otherwise are simple models (though I haven't heard anyone try and see if there might be some set of limitations/optimizations that a technology that simulates the world might be using which would again validate these "additional complexities" for other values)

https://www.smbc-comics.com/comic/2012-02-29
(Not as further proof, but as a example of a similar thought process)

Hide ↑Hide ↑

Harry Maurice Johnston *says:*

> *R is a simpler structure than Z.*

Can you expand on that? It seems to me that, if you start with set theory, it takes quite a bit more work to get to R than it does to get to Z (or, better still, N).

R is more convenient in sometimes surprising ways, but I

wouldn't have said it was simpler.

eyeballfrog *says:*

June 5, 2019 at 11:16 am

@Harry

The first order theory of R is decidable, since every first order statement about R can be algorithmically reduced to a statement about finite collections of open intervals. The first order theory of Z is not, since it contains Peano arithmetic.

tossrock *says:*

June 3, 2019 at 9:20 pm

Sure, but Bostrom's argument is specifically about ancestor simulations, which implies that the simulated world is fundamentally similar to the world doing the simulating.

deltafosb *says:*

June 4, 2019 at 3:54 am

> With respect to available resources and time, we have absolutely no basis for judging whether any such limitations apply within the world running the simulation.

If the universe higher in hierarchy is ruled by some sort of physical laws, we could exploit its thermodynamics of computation (which is pretty universal under general assumptions). If one second of our universe corresponds to N seconds of universe above ours and we a posteriori determine that the reality exists after one second, we've just found a lower bound for the time it takes for the higher universe to reach thermodynamic equilibrium.
That is, if the higher universe can reach thermodynamic equilibrium, respects causality, and you don't believe in Permutation City-esque analytical

continuation of simulations after their shutdown (but in this case we are on our own even if the universe *was* simulated up to first nanosecond after the Big Bang).

Hide ↑Hide ↑

Sok Puppette *says:*

June 4, 2019 at 5:10 am

Least hypothesis is that the simulating universe has the same physical laws as this one. In fact, the whole reason Bostrom gives for expecting anybody to run a lot of simulations is that they're meant to be simulating their own history.

This universe almost certainly does not support running huge numbers of ultra-detailed simulations cheaply, nor is there any real reason to expect that you can skimp on detail in your simulation. Even if it *is* possible, it's *definitely* true that nobody has described an actual way to do it. So it's pretty crazy to say that you're "probably" in one of the simulations. (On edit: and regardless of how much you're willing to spend on simulations, the available computation is in fact finite, no matter how much time you take).

Also, we have no actual understanding of qualia and do not know whether simulations can experience qualia. If you're experiencing qualia, that's at least a little suspicious.

And the **whole damned point of Permutation City** is to riff off the ideas that (a) we don't know what the real substrate is and (b) you can't observe something outside of a simulation from inside it. The point has already been made that it may not even make sense to assign "existence" to anything you can't observe.

Hide ↑Hide ↑

HeelBearCub *says:*

June 4, 2019 at 5:31 am

> *Also, we have no actual understanding of qualia and do not know whether simulations can experience qualia. If you're experiencing qualia, that's at least a little*

> *suspicious.*

I'm not a fan of the simulation argument, but this idea is poor because it assumes its conclusion.

JPNunez *says:*

June 5, 2019 at 4:18 am

The problem is that the nested realities have very little incentive to run their own simulations if they are gonna run super slow.

Besides…look around us! we aren't gonna be running any simulations of the universe any soon. At least not in the scales necessary for another civilization developing inside the simulation.

Ok let's be positive and say that we hit post human status, harness the energy of the galaxy, develop ultra efficient computers, and decide to devote significant amount of resources to simulating a whole another galaxy at 1/1000000 speed, all in a negligible 1000 years in the future.

Then the galaxy stops having useable stars before you hit a good age of the simulated universe to start their own simulation.

Ok, so maybe the whole galaxy starts only a simulation of just a single planet at a more decent speed, let's say 1/1000. Then that simulation has the time to reach our age of the universe, but that simulation does not have the resources to simulate their own, even very slow, galaxy, before the civilzation above dies.

And if we are supposed to be living on a simulation that cannot simulate stuff, well, the argument fails there.

kokotajlod@gmail.com *says:*

June 4, 2019 at 7:52 am

OK... so suppose there's an upper limit as you say. So what? The argument still goes through. The simulation stack can have depth 1 and the argument would still go through.

Randy M *says:*

June 4, 2019 at 8:13 am

But it isn't as convincing if the numbers aren't so overwhleming. It goes from the almost undeniably certain (if stack depth is infinite) to likely but not certain (if simulations are costly). At which point you are much free to be swayed by your intuitions of other considerations.

Loris *says:*

June 4, 2019 at 10:31 am

> There's no (obvious) way of telling how slow (or fast) the simulation is running from within the simulation.

-and-

> OK... so suppose there's an upper limit as you say. So what? The argument still goes through. The simulation stack can have depth 1 and the argument would still go through.

It doesn't, not necessarily. It depends on which version of the simulation argument, or simulation hypothesis we're arguing against.

If what you're saying is that it's possible that the universe we know is a simulation, then yes, it could be simulated. Sure. It doesn't matter that the simulation is running slow.

However, at least some people are considering Nick Bostrom's "simulation argument" trilemma, (copied below from the wikipedia page).

> "1) The fraction of human-level civilizations that reach a posthuman stage (that is, one capable of running high-fidelity

> *ancestor simulations) is very close to zero", or*
>
> *2) "The fraction of posthuman civilizations that are interested in running simulations of their evolutionary history, or variations thereof, is very close to zero", or*
>
> *3) "The fraction of all people with our kind of experiences that are living in a simulation is very close to one"*

The argument for (3) relies on being able to "stack up" simulations. The 'real' universe is large, and will presumably have one or more unsimulated post-human (-equivalent) societies, each with one or more 'universe' simulations – since we haven't accepted (1) or (2).

Each of those universes can then spawn its own civilisations, with their own simulations, and so on. Therefore, most human-equivalents are simulated. Likely including yourself.

But, unfortunately for the argument, it's not possible for the simulations to be able to simulate a larger amount of stuff than they are made of- at least, not in real-time. The 'real' universe is always going to be bigger (and/or more long-lived) than any simulated universes it contains. Even if we accept that simulations can be 'efficient' – that is, the detailed, computationally intensive simulation is limited to a planet surface, say, it's still going to need more than a real planet's worth of stuff. Even if we assume that a posthuman-stage society runs multiple simulations, it is unreasonable to assume that it would dedicate a great deal more resources to that purpose than to propagating itself. So given that we've rejected deep nesting, we can say that even if (1) and (2) are false, (3) could also be false.

This doesn't mean that there will be no simulated human societies, ever, it just means that we're not forced to accept that either we're 'likely' to be in a simulated universe or for civilisation to never reach a 'posthuman' stage.

Hide ↑Hide ↑

tossrock *says:*

June 4, 2019 at 10:43 am

Thank you, I'm glad someone gets it. I think we need terminology for distinguishing these arguments, perhaps the Weak Simulation Hypothesis (P(simulated) > 0) and the Strong Simulation

Hypothesis (P(simulated) ~1, or even > .5). The Weak Simulation Hypothesis is boring, unfalsifiable, and goes back at least to Descartes. The Strong Simulation Hypothesis is bold, interesting, and wrong.

Hide ↑Hide ↑
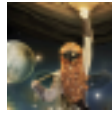
Jameson Quinn *says:*

June 4, 2019 at 11:40 am

There is a level in between these. The Medium Simulation Hypothesis: where P(simulated)>epsilon, where epsilon is some number you could write out decimally without a tremendous/tedious number of zeros – say, the reciprocal of the number of electrons in the sun.

I'd say that WSH is obviously true, but MSH is false, and this latter is actually an interesting fact about reality.

Hide ↑Hide ↑

kokotajlod@gmail.com *says:*

June 5, 2019 at 9:00 am

I was talking about Bostrom's argument as well.

The argument for (3) does NOT rely on being able to "stack up" simulations. In the original paper, Bostrom gives the much better argument that posthuman civilizations would likely create astronomically many simulations (on average), if they create any at all, due to their cheapness:

>A posthuman civilization may eventually build an astronomical number of such computers. We can conclude that the computing power available to a posthuman civilization is sufficient to run a huge number of ancestor-simulations even it allocates only a minute fraction of its resources to that purpose. We can draw this conclusion even while leaving a substantial margin of error in all our estimates.

Note that he is not saying that most posthuman civilizations will create astronomically many simulations. Rather, he is saying that the average number of simulations created per posthuman civilization will be astronomically large, since even if most create zero or one, all it takes is a few to create 10^40 and the average will be astronomical.

Hide ↑Hide ↑
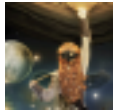
Jameson Quinn *says:*

June 4, 2019 at 11:33 am

My counterargument to simulationism, which seems solid to me, is quantum mechanics. Quantum branching in actual reality means that any deterministic simulation would lead to squillions fewer copies running in sim than in reality. Computer architecture favors deterministic algorithms or at least ones with some deterministic steps. A fully quantum simulation of reality would take at least as many atoms as the reality it was simulating and thus wouldn't really be a "simulation" as we typically think about it.

I think that this argument sets off people's "learned helplessness" alarm bells even more than Bostrom's does — "Quantum multiverse? Probably bullshit." But I could actually do a decent cold explanation of Shor's algorithm (sp? I didn't check because my Wikipedia-free knowledge is the point here); I know what I'm talking about. So I don't think that the fact that nobody smart has engaged with me is evidence that this argument is bad.

(I also think this argument is a lot stronger than the "what do the turtles stand on" argument that launched this subthread.)

Hide ↑Hide ↑

JPNunez *says:*

June 5, 2019 at 4:36 am

I think that then it would be possible to harness the power of the galaxy in our speculative post human state, then run a simulation of a whole universe running at a slower speed…except that universe is fully deterministic, and thus simpler to simulate.

Which would mean the upper simulations would have ultra complex physics, and their simulators decided hahah fuck simulating ultra-quantum mechanics! let's go with just simple quantum physics to sped up the process.
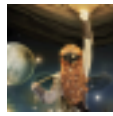
Jameson Quinn *says:*

June 5, 2019 at 5:12 am

One trivial point, and two counterarguments that I think are independent and individually sufficient:

Trivially, I think it's better to structure the spatial metaphor with the outer simulations on the bottom, not the top. Turtles, substrates, roots, etc.

Counterargument one: anthropically, we can expect to live in the most insanely multiple physics that exists (and thus, if you think existence itself is merely a matter of logical possibility a la Tegmark, that's logically possible), and that can support consciousness. [Note I say "multiple" rather than "complex"; I'd also accept "creative" or "productive".] So if our quantum universe is just a sim running in a hyper-quantum universe, most consciousness would be in the hyper level.

Counterargument two: even if there exists a hyper-universe capable of trivially simulating a quantum universe/solar system/planet, it would be essentially impossible to intervene (perform miracles) in more than an infinitesimal fraction of the Everett branches of the sim, so for most people inside the sim, there would be literally no difference between their reality and actual reality. If the Tegmark logical multiverse is correct, then, most of the sim would merely be reproducing a universe that already existed independently in un-simmed form. So most people inside that sim could correctly say that they were not in a sim.

I think that counterargument 1 is substantially stronger than 2; though confidence levels are a bit arbitrary, I'd give on the order of

99.9% confidence to #1 and 10% to #2, but since I couldn't concretely imagine any alternatives to #2 I say I "believe" it anyway. If either one is true, Bostrom's argument is wrong.

JPNunez *says:*

Counterargument 1 is brutal. Simulations are going to be necessarily less efficient at supporting conciousness than, you know, actual real people living. Way more if you need to run the simulated universe at 1/100 of speed to get any decent size/precision.

This fails with "Age of Em" style simulations, but those know they are artificial, because they are not running a simulation of physics, just a simulation of people.

So chances are you are real.

Counterargument #2 is weird, because it just argues that simulations count as real? Dunno if miracles are needed to argue that a universe is simulated.

Tegmark argument I am not so sure; I am only vaguely familiar with it, but IIRC Tegmark would argue that conciousness is a mathematical construct, thus you are real whether you live in a simulation or not.

Which is a moral argument that I feel it is independent of Tegmark?

Someone upthread mentioned that the rules of mathematics may be different in the realer, down-turtle universes, but I feel that is impossible, particularly if simulations themselves are mathematical constructs.

I tend to think as simulations as "down". Maaaybe based

on the Inception movie and its "we need to go deeper" phrase.

But it probably makes sense both ways. Up turtle for simulations, down turtle for universes that simulate others.
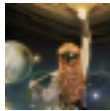
Jameson Quinn *says:*

Note that counterargument 1 also implies/predicts that QM is among the most insanely multiple possible substrates for consciousness. Which I find plausible; and I consider that plausibility to be weak evidence in favor of Tegmark.

JPNunez *says:*

You mean because the more complex the substrate, the higher amount of sentient beings it can support? Not sure I follow.

kokotajlod@gmail.com *says:*

You get crazy results if you proportion credence according to number of branches. (In fact I'm not sure it even makes sense to do so?) Instead, the standard view is that you should proportion credence according to the Born rule, which uses the amplitude of the branches.

So if you have a simulated system that is mimicking a real system, and the real system is constantly branching, and the simulated system isn't… then it shouldn't make a difference to how likely you think you are to be in one vs. the other.

(Caveat: I'm told it is really the square of the amplitude or something like

that. So actually this might advantage the simulation over the non-simulation!)

Hide ↑Hide ↑
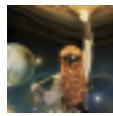
Jameson Quinn *says:*

June 9, 2019 at 11:57 am

The Born rule is also crazy. It means that your measure is fading exponentially by the second.

I've seen and, I thought, understood other alternatives, but cannot currently find or recreate them. Certainly I think that "naïvely counting branches" and "Born rule" are not obviously the only options.

Hide ↑Hide ↑

Peter Gerdes *says:*

June 3, 2019 at 7:47 pm

As far as Pascal's mugging goes* there is a relatively easy response. If you really were logically omniscient and utility effects do increase faster than probabilities decrease then it very much does make sense to focus efforts on very unlikely outcomes with huge utility effects. I think it very much does make sense to spend a huge effort preventing the unimaginable horror of BusyBeaver(2^^^999999) people spending that many years in horrific torture.

What isn't justified is focusing on outcomes where you have no idea if your actions are an expected utility positive. For instance, in cases like the actual pascal's mugging you have no idea that caving to the threat makes that outcome less, not more, likely. Moreover, you have every reason to believe that your beliefs about such cases are likely to be subject to ineliminable intrinsic bias so even when you have a strong intuition (like some people do about complying with threat) that some action reduces these super unlikely but high impact events you can't trust it.

In other words where things go wrong is in the implausible assumption that the **idealization** we make when we model people as having priors continues to hold even when we consider super bizarre and unlikely outcomes. In actuality people have no way to comparing the probabilities of these outcomes conditional on various acts of theirs.

Or to put the point more formally it's most appropriate to just model us as not assigning probabilities to some events at all. These include the pascalian mugging situations so we can't even compute that the likely effect of an intervention in these cases is positive.

Bugmaster *says:*

IMO Pascal's Mugging is a good illustration of why talking about "superintelligences" or gods is kind of pointless. Once you start introducing infinite (or practically infinite) terms into your equations, your math breaks down and you can't really calculate anything useful, one way or the other.

kokotajlod@gmail.com *says:*

If you are interested, I have a [sequence](#) on this topic.

TL;DR: If you really were logically omniscient etc. etc. then you would be paralyzed, since every action would have undefined expected utility. If you are not logically omniscient but try to approximate it by considering more and more possibilities, your behavior will become increasingly erratic in the limit as you approximate it better and better. Hence it is silly to hold up expected utility maximization as an ideal to guide our reasoning.

…unless you have a weird, restricted prior, or a bounded utility function (I do!).

carsonmcneil *says:*

You've never heard a good argument against Pascal's Wager/Pascal's Mugging?!? Those guys are pushovers! (Though in a sense, Pascal's Mugging IS a weak-ish argument against Pascal's Wager)

The fundamental issue with Pascal's Wager is not that the cells of the payoff matrix are wrong, it is that the matrix is woefully incomplete. It assumes that you have two choices: believe in a

nice sane Christian God that gives you all the rules to get into heaven, or be an atheist. This is obviously a totally false premise. If you fill out the matrix with the complete space of possibilities, you not only have to add all the other (mutually exclusive) world religions, but also the possibilities that I thought of right now while writing this: such as inverted-Yaweh, the one true God that will send you to hell for being a good Christian/Jew.

If you accept that some of the cells in Pascal's Wager are infinity (presumably because hell is torture forever, not because it is an infinite amount of torture every moment), the result of actually including all the possibilities is that you have infinities and negative infinities everywhere, and you can't compare them. Infinity-infinity=undefined. Also, infinity*small-probability – infinity*large_probability=undefined, so the already-flawed argument that my made-up anti-Yaweh is less likely than Yaweh does not matter. If you don't accept that the payoffs are infinity (or negative infinity), then you must provide probabilities and we're back to square one.

To actually compare the infinite space of hypotheses about potential dieties or non-dieties, you need considerably more advanced math from statistical learning theory. You need to talk about the VC dimension of different hypothesis spaces. Needless to say, doing so does not yield a conclusion anything like Pascal's Wager.

(Also as an aside, I am probably one of these people that Scott says is susceptible to radicalization, as I don't feel like I've ever heard an argument that prescribes I take some radical action which I am not that I couldn't find some flaw in that at least wrecks the rigor of the argument in some difficult-to-evaluate way.)

HideHide

kokotajlod@gmail.com *says:*

June 4, 2019 at 8:05 am

If we interpret the arguments literally, as "You should believe in God" or "You should give me the money" then yes they are pretty easily responded to.

If we interpret them as open problems in decision theory / rational choice that make a mockery of our ideal of expected utility maximization, then they are pretty difficult to respond to. Because, as you say, you quickly realize that every action has undefined expected utility…

I'm interested to hear more about this VC dimension stuff. Does it solve the problem mentioned above? Do you have anything written about this I could read?

DavidFriedman *says:*

June 3, 2019 at 8:20 pm

> *or admit even the slightest possibility that you might be right.*
>
> *(This is the correct Bayesian action: if I know that a false argument sounds just as convincing as a true argument, argument convincingness provides no evidence either way. I should ignore it and stick with my prior.)*

I disagree. Some false arguments are convincing, some true arguments are not. But being true is an advantage in convincing people, just not an overwhelming advantage. If that were not the case, then nobody, including the experts, could ever learn anything.

Once you have heard a convincing argument for something you didn't believe, you should admit the possibility that it is right–raise your subjective probability. Just not very high.

Matt M *says:*

June 4, 2019 at 8:07 am

> *But being true is an advantage in convincing people, just not an overwhelming advantage.*

Only in cases where truth can be absolutely established. And I think part of the point here is that such cases are few and far between. At least, among the set of "things people are likely to argue about."

Randy M *says:*

June 4, 2019 at 8:16 am

I'd say actually in all cases where some of the truth can be established. A true story is going to be internally and externally coherent in ways that false stories may not be.

But the best lies are mostly true, possibly containing as much 'known truths' as the competing wholly true theory.

Matt M *says:*

I'm not certain that's right. Someone above pointed out that liars have "more degrees of freedom" which I would say is correct. The actual truth sometimes doesn't "fit together" as well as a carefully assembled lie.

See also: Concepts like "too good to be true" or "you can't script that." When really unlikely things happen in say, a sporting event… is that evidence that sporting outcomes are genuine because nobody would try and "fake" something so dramatic and implausible? Or is it evidence that sporting outcomes are rigged because left to random chance, we wouldn't get such dramatically implausible results?

AlexOfUrals *says:*

In theory, yes. But in practice, you have a finite amount of resources to spend on any given task, and so you evaluate it with finite precision. If you estimate that the resulting change from your evaluation of the argument will be below that precision level, you might just as well decide not to evaluate it at all and save some resources. Moreover, evaluating the argument and then correcting on the fact that you're easily convinced by a false argument is essentially a task of separating signal from noise, and your resources and abilities at that are also limited. If you know you can't do it effectively, you'll be adding noise to your judgement, thus making it less precise, not more, even as you take into account more information. In this case refusing to even consider the argument not only saves you resources but also leaves you with better (more trustworthy) knowledge on the subject.

I'm not sure that's exactly what Scott meant in this quote though, but that's my rationale behind a roughly similar approach to many topics.

DavidFriedman *says:*

June 3, 2019 at 8:24 pm

Scott's story about Velikovsky parallels an experience I had when I was looking at colleges. Yale happened to be having a program on the controversy over HUAC, the House Unamerican Activities Committee. The first part was a movie, Operation Abolition, which convincingly demonstrated that the campaign to abolish HUAC was a communist plot. The second was another movie whose name I have forgotten, which convincingly demonstrated that the first movie was a fraud. Then there was written material debunking the second movie, and more written material on the other side.

My conclusion was not Scott's, that you should never let yourself be persuaded in such a context. It was that you needed to hear both sides of the argument before forming an opinion.

And sometimes, of course, it should not be too confident an opinion.

HideHide

Enkidum *says:*

June 3, 2019 at 8:42 pm

But it's not that you should never let yourself be persuaded in such a context. It's that there are thousands of such contexts, and learning enough to be a good judge of any of them takes years (and even then is far from foolproof). The best of us might manage to be good judges of a dozen or so over the course of our lives, perhaps two dozen. What should we do, then, in the vast majority of cases where we will not become experts?

EDIT: Which is not to say that the answer to my question is "form no opinions in those cases". But it is to say that "trust experts even if you don't invest the time to understand them" is, all things being equal, a pretty good strategy much of the time.

Hide ↑Hide ↑

Tibor *says:*

June 4, 2019 at 5:49 am

To me Scott's argument sounds more or less like that of rational ignorance. Regardless of what opinion you form about ancient history, it won't fundamentally affect your life enough to make the effort necessary to form a good opinion worth it. If you know to

know you don't know enough and if you want to pride yourself on your rationality (he lesswrong flavour of rationality) the best course of action is not to form an opinion at all.

Of course, in practice people form opinions anyway, either because the opinion doesn't really matter and it is more fun than not to have one (could be pragmatic too if it means you fit in better with your peer group and people generally want to fit in since that comes with a lot of practical advantages) or because they don't know they don't know enough.

But I'd say that truly not having an opinion is harder than Scott makes it. On a very intellectual level that might as well be the case, but if one estimates true opinions (if there is such a thing) by actions rather than words, it is sometimes impossible not to have an opinion on a subject (ancient Mesopotamian history is arguably usually not the case).

Hide ↑Hide ↑

Michael Handy *says:*

June 3, 2019 at 8:51 pm

This would seem to imply that the amount you should listen to an argument should be the inverse to the persons intelligence and skill in rhetoric.

That is. If your idiot friend is stuttering his way through an argument and it seems convincing despite his total inability to string two coherent chains of thought together, it is much less likely that its a brilliant, clever and wrong argument hacking your brain.

HideHide

sty_silver *says:*

June 4, 2019 at 2:40 am

I think there's something to that. ImE it's very rare that stupid people say things that sound convincing, but if it does happen, you probably ought to take it seriously.

Hide ↑Hide ↑

Matt M *says:*

June 4, 2019 at 8:13 am

Agree with this. Maybe less with the intelligence part, and more with the "skilled at public speaking" part.

If an argument from an awkward dude who sucks at public speaking sounds just as effective as an argument from a professional lawyer or motivational speaker or politician, I'm going to place a large prior on the awkward dude having the correct position, all else equal…

Hide ↑Hide ↑

[Jaskologist](#) *says:*

June 4, 2019 at 11:44 am

Which is why lawyers and politicians will often adopt the persona of a [Simple Country Lawyer from the backwoods.](#)

Hide ↑Hide ↑

Shion Arita *says:*

June 3, 2019 at 9:51 pm

To follow a certain vein, no good counterartuments to antrhopic doomsday?

The whole thing is incoherent from the start. There is no 'probability that I was born as someone else'. Given the fact that I exist as myself, I was necessarily born as myself. And the same goes for everyone else. Counting people is not like counting electrons; what we call 'person' is merely a label for similarities that exist in certain patterns that exist in the universe. It is not a label for actually identical patterns. If we couldn't distinguish one person from another, then maybe the argument would mean something, but since we can, it doesn't. Probability is about our knowledge, not anything fundamental in this case. We are not asking "Alice and Bob voted in the election. Someone voted for Trump and someone voted for Clinton. What is the probability that Alice voted for Clinton?". We rather are asking in this case, "Alice and Bob voted in the election. Alice voted for Clinton and Bob voted for Trump. What is the probability that Alice voted for Clinton?"

What is the probability that I was born as me? 1.

HideHide

kokotajlod@gmail.com *says:*

June 4, 2019 at 8:20 am

I disagree: The whole thing is pretty darn coherent. It's just ordinary, vanilla Bayesianism applied to your whole life's worth of evidence instead of just one particular update. I'd be happy to explain more if you like.

Unless you are saying that there is no such thing as self-locating belief? If an experimenter duplicated you with some sci-fi technology, your credence that you were the original vs. the duplicate would be undefined? Maybe you are advocating for Full Non-Indexical Conditioning (FNC) or Meacham's Compartmentalized Conditionalization? Those theories I classify as non-vanilla Bayesianism, and they do indeed avoid the argument, but they have much bigger problems as a result.

Unless unless you are going for Anthropic Decision Theory / Updateless Decision Theory? I think this is plausible actually.

Hide ↑Hide ↑

ksvanhorn *says:*

June 4, 2019 at 12:02 pm

> *I disagree: The whole thing [anthropic doomsday] is pretty darn coherent.*

I disagree with your disagreement :-). The anthropic doomsday argument is flawed, in that it makes an assumption that violates probability theory. Here's a careful analysis:

Doomsday and the Dice Room Murders

Hide ↑Hide ↑

kokotajlod@gmail.com *says:*

June 5, 2019 at 9:53 am

Thanks for sharing this; I hadn't seen it before and it is an interesting and serious attempt to grapple with the problem.

I think it is wrong though. I haven't read through the math yet (if you challenge me to do so, I will find time to do so 🙂) but here's a visual demonstration of the possibility of what the post claims to be

impossible:

You have an infinite grid of infinitely deep boxes. You have a finite amount of water to spread around.

At (1,1) put a drop.
At (2,1) put 35/36th/11 of a drop.
At (2,2) put 35/36th/11 of 10 drops.
At (3,1) put (35/36^2)/111 of a drop.
At (3,2) put (35/36^2)/111 of 10 drops.
At (3,3) put (35/36^2)/111 of 100 drops.
At (4,1) put (35/36^3)/1111 of a drop.
At (4,2) put (35/36^3)/1111 of 10 drops.
At (4,3) put (35/36^3)/1111 of 100 drops.
At (4,4) put (35/36^3)/1111 of 1000 drops.
… (continue this pattern infinitely)

If I've set it up right, the total amount of water used will be finite, even though you use infinitely many boxes.

The water is a metaphor for your total credence, which sums to 1. Each drop (or partial drop) is a possible person. The first row is the possible world in which the madman gets one person, rolls snake-eyes, and kills them. The second row is the possible world in which the first person survives but the second batch, of 10 people, rolls snake-eyes and dies. Etc.

Your total credence in each world is actually diminishing as we travel to higher and higher rows–representing the fact that each world is 1/36 less likely than the previous world due to the additional roll of non-snake-eyes needed to bring it about.

However, in each world/row, your total credence is ~90% in the farthest box, the box in which most of the drops/people are–the box of doomed people.

So ~90% of your total credence/water is in doomed people. So you

are probably doomed, since you don't have any other information that lets you rule out being some of the people.

This is, I think, a model of how we should analyze the situation, and it supports the view that the probability of death is ~90%. I've commented on the original post to see what the author thinks.

Edit: In this model I was assuming there are no other people in the world besides the madman's victims. If instead we think the world has some large finite population N from which the victims are drawn at random, then the credence in each world/row should be multiplied by its number of victims–so overall you will think you each new row is more likely than the last–but we don't have infinitely many worlds/rows, we only have finitely many, so your credence can still sum to 1.

Hmm, I notice I am confused–for I still haven't dealt with the case in which we combine a larger population N from which the victims are drawn with infinitely many possible values of N. I'll have to think about this more to see if that changes anything.

Hide ↑Hide ↑

uau *says:*

June 5, 2019 at 10:19 am

I think this is closely analogous to escalating bets. Suppose that in a game you either win or lose the same amount, and have a 30% chance of winning. You can place bets for arbitrary amounts. Is it a winning strategy to keep multiplying your bet by 10 each time you lose? Most money will be in the bet you win…

Hide ↑Hide ↑

uau *says:*

June 5, 2019 at 10:57 am

I think this might be a reasonable overall explanation of what is happening here:

Fix some order in which people will be selected for groups (exactly who will be in the group of 1000 if it occurs in the game, and so on). Now keep iterating the game. For any particular person, the following holds true:

(number of games in which this person has died) / (number of games in which this person has been selected) will go to 1/36 as the number of games approaches infinity.

HOWEVER, due to the exponential/divergent nature of the game, after any particular finite number of games, the group of people who have been selected at least once so far is dominated by people who have not approached the true limit yet and have died more.

Hide ↑Hide ↑

ksvanhorn *says:*

June 6, 2019 at 12:08 pm

You need to read the post in detail. If you assume a finite population, then there's a possibility that the madman runs out of people to kidnap and ends up never killing anybody; accounting for that drastically changes the result of the analysis.

That scenario doesn't work well as a metaphor for extinction risk, so assume an unbounded (potential) population. The Doomsday Argument relies on an argument that *pi(t+1)*, your prior probability of belonging to generation *t+1*, is 10 times larger than *pi(t)*, your prior probability of belonging to generation *t*, and that this holds for **all** *t*. That simply can't be true; there is **no** proper probability distribution over the positive integers having that property. At some point the prior probabilities have to start **decreasing** in order to sum to the finite value 1.

Hide ↑Hide ↑

uau *says:*

> *You need to read the post in detail.*

Why would you say that? I mean, I think I explained what's going on in the madman scenario. You seem to imply I missed something, but nothing you then follow with really looks like that.

Hide ↑Hide ↑

wanda_tinasky *says:*

June 4, 2019 at 3:46 pm

I sharply disagree. The DA is trivially false

a) We're not a random sampling from the cohort of all humans. (This fact is completely dispositive, btw. I'm going to continue but any rebuttal that doesn't clearly address this is vacuous.) At best we're an 'ordinal sample', which provides no information about future samples.

b) Even if you wanted to posit some hypothetical Cohort of All Humans, we aren't born with serial numbers on our head, so there's no way to infer any statistics about the parent population. To put it in terms of the German Tank Analogy: imagine no serial numbers on the tank; how are you going to estimate the number German Tanks?

The only thing the Doomsday Argument provides evidence for is the fraudulence of modern academic philosophy.

Hide ↑Hide ↑

kokotajlod@gmail.com *says:*

June 5, 2019 at 9:15 am

re:a: True, but irrelevant– see Anatid below for an explanation of why this issue is a red herring; the DA doesn't depend on it. Like I said, all you need is vanilla Bayesianism: Just think about making

the update from not knowing your birth rank to knowing your birth rank.

re: b: We aren't born with serial numbers on our head, but we have a pretty good idea of what our birth rank is. Imagine the German Tank Problem but with only 1 tank, and the serial number has been filed off but you can see how many digits long it was.

re: the fraudulence of academic philosophy: I have my own gripes about academic philosophy, but this is not one of them. Philososphy, unlike say physics, is cursed, in that people who think themselves smart are likely to think themselves philosophically sophisticated also. Non-physicists who talk about how shitty mainstream physics is are generally recognized as cranks, but non-philosophers who say the same about philosophy are too numerous to be swatted down.

Hide ↑Hide ↑

---

wanda_tinasky *says:*

June 5, 2019 at 1:23 pm

I'm sorry, but a) absolutely is relevant. There is no updating that happens upon learning the birth order.

>Imagine the German Tank Problem but with only 1 tank, and the serial number has been filed off but you can see how many digits long it was.

I mean, that just completely misunderstands the objection. German tanks carry information about their parent population in the form of serial numbers, but only because they're being observed in a randomized fashion! I mean, imagine they were always encountered in exactly the order that they came off the assembly line. That's the equivalent of numbering the gumballs after they come out of the machine. It provides zero information about the parent population.

The curse that philosophy has over physics is that the

language it uses isn't precise enough to mechanically prove that the nonsense it occasionally tolerates is, in fact, nonsense. This fact gives endless cover to the psuedo-intellectuals who make profitable use of Brandolini's Law (aka The Bullshit Asymmetry Principle).

Hide ↑Hide ↑

Anatid *says:*

June 4, 2019 at 8:17 pm

> *There is no 'probability that I was born as someone else'.*

I don't think this concept is necessary for the doomsday argument.

Pretend all humans are fully conscious and intelligent in the womb. Five minutes before each human is born, before they know anything about the world, a being from the end of time appears them in a vision and offers a wager. You can bet $100 that you are in the last 50% of humans to be born. When the being appears to you, what odds are you willing to accept for this bet?

I claim that you should demand at least even odds, since exactly half of all humans will win this bet.

Hide ↑Hide ↑

wanda_tinasky *says:*

June 5, 2019 at 3:04 pm

All you're saying is that there's always a 50% probability of being in one of two equally likely pots, which is an empty tautology. This has no bearing whatsoever on the Doomsday Argument.

Hide ↑Hide ↑

Anatid *says:*

June 5, 2019 at 8:25 pm

Sure, it's basically a tautology. Sounds like you accept that before seeing the world, you should believe that you have a 50% chance of being in the last half of humans. Now, the being from the end of

time tells you one more thing: 100 billion humans were born before you.

Does learning this one fact change your probability that you are in the last half of all humans? I think it shouldn't; it doesn't seem like the kind of information that can change the probability on its own.

So, from your position of ignorance inside the womb, shouldn't you assign a 50% probability to the claim that no more than 100 billion humans will be born after you? And, generalizing, a 90% probability that no more than 900 billion humans will be born after you?

I think so. As nadbor discusses below, if you have more information that gives you some real ability to predict the future course of human history, then you can improve on the naive doomsday probabilities, but I think they are the correct baseline if your starting information is just how many humans have lived so far.

Hide ↑Hide ↑

wanda_tinasky *says:*

June 6, 2019 at 6:50 pm

Taking 9-1 odds on a 10% likely proposition is tautologically correct. That provides no information. The strategy is the same regardless of how distant the doomsday is, and therefore it definitionally has no predictive power.

I'll also point out a subtle error that is introduced by framing it as a conversation with a Being Outside Time: it makes oneself appear to be a random sample from the timeline of all humans. This is purely a framing effect illusion. One isn't randomly selected from a pre-existing unborn population and then observed as having a birth order (that *would* be a valid way to infer statistics about the unborn population – unfortunately, we can't sample from the future). The framing effect becomes clear when

you replace "being outside time" with "human archeologist who for some reason hangs out in hospitals telling babies what their birth order is". When that mathematically-irrelevant-but-emotionally-meaningful substitution is made, it's obvious that the archeologist's answer can only depend on the past, and no information which does not depend on the future can yield any bayesian update about the future. Which, of course, should be obvious for multiple other reasons, not least of which is that information doesn't travel backwards in time.

Hide ↑Hide ↑

Anatid *says:*

June 6, 2019 at 8:49 pm

If we are able to say things like, it's only 10% likely that more than 900 billion humans will be born after us (from the position of ignorance inside the womb, knowing only how many people have been born so far), that sure sounds to me like we have some predictive power? Even if it's tautologically true. It's definitely something we wouldn't have been able to say without knowing that 100 billion humans had lived so far. So it feels to me like we learned something about the probability distribution over the number of future humans by learning the number of previous humans. But maybe this is just an argument over what "predictive power" means.

If you don't like the wager being decided by a Being Outside Time, we can switch it up to respect causality: after humanity goes extinct, and the outcome of the wager is determined, all humans are resurrected in heaven and the outcome of the wager affects their initial heavenly bank account balance. (I'm just using the wager to elicit honest probabilities from the baby, who needs to see some reward for a correct guess).

Here's a "doomsday argument" that tries to avoid even the

appearance of implying that one is randomly selected from some overall population. Google tells me there have been 66 British monarchs so far. If I tell this to someone who doesn't know what Britain is, and has never heard the word "monarch", would it be reasonable for them to conclude that whatever British monarchs are, there will probably never be as many as 1 million of them total?

Hide ↑Hide ↑

JoeCool *says:*

June 3, 2019 at 11:51 pm

This is kind of why I think democracy is very suboptimal.

Basically, we make voting rights and political opinions this symbol of valuing someone as a human being, yet almost everybody can't really be bothered (and as this post says, shouldn't be bothered) to have coherent/even slightly informed/not systematically misinformed opinions, but we can't really politely say this out loud.

I mean the Washington post tagline is democracy dies in darkness, which is funny because objectively speaking the electorate has and always will be in darkness.

Then again, its very difficult to argue with the results of democracy.

Freedom of speech in particular seems like a very valuable societal self checking mechanism, at least up to a point, and it never seems to exist without democracy.

HideHide

Aapje *says:*

June 4, 2019 at 5:50 am

That's why most countries don't have direct democracy. We don't choose policies, we choose our political masters.

Hide ↑Hide ↑

JPNunez *says:*

June 5, 2019 at 5:10 am

I assumed that we don't have direct democracy because it'd be super impractical. Imagine having to hold a country wide vote every week on whatever the equivalent parliament-we-have-today would be discussing. It would be very disruptive to the normal work of the country, and expensive.

I guess we could all do it electronically in the future. Then the whole "the electorate is not well informed" argument would be relevant.

checks wikipedia uh switzerland has something like that.

Let's discount it as random error.

Hide ↑Hide ↑

Aapje *says:*

June 5, 2019 at 6:38 am

The Swiss still have a parliament that propose and pass laws. The populace can reject the laws, so it's more like how the courts in the US can overturn passed laws, but then with that power in the hands of the people.

So it's more of a mixed system, than actual direct democracy. This is why it's also referred to as half-direct or representative direct democracy.

Hide ↑Hide ↑

JoeCool *says:*

June 7, 2019 at 5:29 pm

The problem of ignorance still exists to a large degree with representative democracy.

The only thing saving us is the representatives themselves knowing that if they give the people what they want it would hurt the economy and people would vote them out, so its this giant manipulation game.

Hide ↑Hide ↑

DavidFriedman *says:*

Except that most things that hurt the economy take long enough to visible do so that the voters don't know who to blame.

Hide ↑Hide ↑

Clutzy *says:*

June 7, 2019 at 10:33 pm

Not at the local level, where something like an actual socialistic city would collapse very quickly.

That is why socialists are typically also in favor of systems having sovereignty over many people, the system collapses extremely quickly if people can foot vote out of it.

Hide ↑Hide ↑

Snickering Citadel *says:*

June 4, 2019 at 1:12 am

The doomsday argument seems like nonsense, but I don't understand all the math, so maybe not. I'll try to argue against it anyway.

The doomsday argument seem to say first there is the history of all humanity. Then there is bunch of consciousnesses that get distributed randomly among the humans, one consciousness for each human. Since I have a random consciousness I can make predictions on where in the line of consciousnesses I am.

But I don't think I have a random consciousness. When the first human appeared, the first human consciousness appeared. Not a random one, but the first one. When the second human appeared the second consciousness appeared. So you can't deduce the total number of humans that will ever appear. That number is not decided until all humans have appeared.

Also it seems like the doomsday argument gives us information of the future. Could that not result in a grandfather paradox? Like we could postulate universe with a person that has the ability to destroy all of humanity. And he has decided that if the doomsday argument says he is in a universe where humanity will die out soon he will NOT destroy all of humanity. And if he is

a universe where the doomsday argument says humanity will NOT die die out soon, he will destroy all of humanity.

HideHide

Bugmaster *says:*

June 4, 2019 at 3:13 am

I confess, I also don't understand the Doomsday Argument. Is it trying to predict, based on the current and historical number of living humans, how long humanity will exist ? Isn't that kind of like predicting the weather based on how many eggs you ate for breakfast this morning ? I mean, yes, surely there's some correlation, but why would anyone pick such a circuitous prediction method ?

Hide ↑Hide ↑

LadyJane *says:*

June 4, 2019 at 3:26 am

@Snickering Citadel, @Bugmaster: It's basically just the same logic behind the German Tank Problem. If you see a tank with the serial number 25, then you could reasonably guess that the total number of tanks is in the dozens or hundreds, maybe the low thousands at most. Sure, it's possible that the tank is actually 25 out of 999999, but that's significantly less likely. Likewise, it's possible that humanity will go extinct tomorrow, or that humanity won't go extinct until matter itself stops existing in a trillion trillion years from now. But those are significantly less likely outcomes than the range predicted by the Doomsday Argument.

Hide ↑Hide ↑

Bugmaster *says:*

June 4, 2019 at 3:34 am

I've never heard of the German Tank Problem, but I confess, I don't understand it either. How do you know there isn't a factory somewhere, churning out tanks at the rate of 5 tanks/day or something ?

Hide ↑Hide ↑

Aapje *says:*

@LadyJane

> *It's basically just the same logic behind the German Tank Problem. If you see a tank with the serial number 25, then you could reasonably guess that the total number of tanks is in the dozens or hundreds, maybe the low thousands at most.*

In the German Tank Problem, you want to know how many tanks *have already been produced & fielded* and you try to figure this out by looking at the serial number of a random, already produced tank.

It's unlikely that this will be a serial number low in the range and it takes only a few samplings to get a fairly good idea of the plausible number of produced tanks, assuming that the selection is truly random (which is actually quite unlikely) and assuming no countermeasures. Producers of military equipment actually quite commonly fudge the serial numbers, rather than use monotonously increasing numbers, starting with 1 (although this is irrelevant for the thought experiment).

You can't use the estimate of already produced tanks to figure out how many tanks will be produced *in the future* and especially not when production is going to end, just like you can't use a single speed measurement to measure acceleration or to figure out when the car is going to crash.

You could use the observed increase in serial numbers over time to establish a trend line, just like you can make a trend line based on population growth in the past. However, that tells you nothing about black swan/doomsday events that cause the production of German tanks to suddenly stop or that wipe out humanity.

The doomsday argument deceives people by misrepresenting the situation, by framing it as if all people already exist and you take a random sampling from all humans. In actuality, the sampling is the

opposite of random, as well as being restricted to already existing humans: it always looks at the most recently born human. The only conclusions you can draw from this are completely trivial and don't tell us anything about the future.

Just like the Monty Hall problem, the doomsday argument takes advantage of humans being naturally rather poor at statistics, having heuristics that only work well for fairly simple problems. You can get people to apply the wrong heuristics for a situation by disguising a problem like a different problem.

Hide ↑Hide ↑

Aapje *says:*

June 4, 2019 at 7:11 am

Let me give an example to illustrate:

You notice a robot walking around with serial number 25. A reasonably guess is that there are not that many robots walking around, with 50 being the most likely number.

However, you can't conclude that only 50 robots will ever exist. Perhaps the robots have been on sale for only one hour and they actually sell at 50 units per hour. Then if this rate holds, there soon will be thousands of robots walking around and eventually there will be millions, if the robots keep selling. However, the factory may also be continuously be ramping up production. In that case, to predict the number of robots, you need to know both the production rate and the production rate increase, to predict the future.

Just knowing the current number of robots is in itself far from sufficient to predict the future.

—

how long they are being sold. Some products like Coca Cola, last for centuries, while others get replaced quickly, even if they briefly had huge sales.

Then the doomsday prediction for products is presumably best guessed by applying historic data to the products. An already long lasting product is likely to keep lasting, while newer products are more likely to stop being sold. It's very unlikely that Coca Cola will disappear from shops tomorrow, for example.

So then humans, who are already immensely successful, are like Coca Cola, with the caveat that some moron may try to introduce New Coke.

😉

Hide ↑Hide ↑

LadyJane *says:*

June 4, 2019 at 1:19 pm

> *Imagine that you noticed that products differ greatly in how long they are being sold. Some products like Coca Cola, last for centuries, while others get replaced quickly, even if they briefly had huge sales.*
>
> *Then the doomsday prediction for products is presumably best guessed by applying historic data to the products. An already long lasting product is likely to keep lasting, while newer products are more likely to stop being sold. It's very unlikely that Coca Cola will disappear from shops tomorrow, for example.*
>
> *So then humans, who are already immensely successful, are like Coca Cola, with the caveat*

> *that some moron may try to introduce New Coke.*

That's actually consistent with the Doomsday Argument! As mentioned below, if the Doomsday Argument was applied back when humans were still a new species, it would've predicted that humanity wouldn't be around for much longer, just as one might predict that a new restaurant is likely to go out of business. That prediction would've been *wrong*, but that's okay; by definition, 5% of the humans in existence (past, present, and future) will be outliers who'd be wrong if they tried to apply the Doomsday Argument to themselves.

Keep in mind, the extinction range produced by the Doomsday Argument is enormous: It predicts that humanity has a 97.5% chance of going extinct at some point between 5,100 and 7,800,000 years from now. Some people are only familiar with the variant arguments that give much more specific figures, but those are "impure" versions of the argument that attempt to take various other factors into account. For instance, the most commonly known DA variant predicts extinction in 9,120 years, but it was calculated taking expected population growth rates into account, which makes it significantly less likely to be true.

Hide ↑Hide ↑

Aapje *says:*

June 4, 2019 at 3:43 pm

@LadyJane

Part of the uncertainty is that we are probably more different from other animals than Dr. Pepper is from Coca Cola. So we might be abnormal in [either](#) [direction](#).

We might be combo-breakers.

Ultimately, it's highly uncertain, not just how long we last, but simply how fragile we actually are.

sty_silver *says:*

June 4, 2019 at 3:46 am

The argument is this: if humanity goes extinct at some point, then the number of all humans who've lived is finite. Thus we can put them in one ordered list (ordered based on their time or birth), i.e. (h_1, …, h_n). You are one of those people, say h_k, where k is between 1 and n.

Now, the probability that you appear in the first tenth of the list is 1/10, because, well, it's a tenth of the elements. Similarly, the probability that you appear in the first millionth of the list is 1/1000000.

If humanity goes exinct in 50 years, you'd be roughly in the middle (about a tenth of all humans who have ever existed exist right now). If it lasts for another 1000 000 years, you'd be in the first millionth. The probability for this is exceedingly small, therefore the premise is exceedingly unlikely. Thus, it's far more likely that humanity goes extinct in 50 years, because then your position among the list is typical.

Bugmaster *says:*

June 4, 2019 at 4:30 am

> *The probability for this is exceedingly small*

Why ? Aren't you assuming a fairly uniform distribution of humans over time ? Or, to put it another way: if you rewind time a little, and apply this argument to humans living in the first great city of Ur, wouldn't you doomsday prediction fall spectacularly short of the mark ?

sty_silver *says:*

June 4, 2019 at 4:53 am

I think the DA is wrong, so I won't try to justify it, I was just explaining what it actually argues for. I don't really think LJ's summary was accurate.

Hide ↑Hide ↑

Bugmaster *says:*

June 4, 2019 at 6:09 am

@sty_silver:
Fair enough — I appreciate the explanation.

Hide ↑Hide ↑

nadbor *says:*

June 4, 2019 at 9:07 am

Yes it would and it would be wrong. But that's to be expected. By definition DA estimate is correct for 95% of humans who use it and wrong for 5%. People in Ur were among the 5% for whom it is wrong. We may or may not also be among the 5%.

Hard to tell what we can do with this information, if anything. See my response to Furslid below.

Hide ↑Hide ↑

Watchman *says:*

June 4, 2019 at 9:53 am

Doesn't the domesday argument require a human population that is growing or stable. If human numbers are on a normal distribution, a reasonable conjecture considering we currently have declining birth rates globally and are seeing population decline set in in some countries, then won't the assumptions have to change?

Hide ↑Hide ↑

Furslid *says:*

June 4, 2019 at 4:41 am

My problem with the doomsday argument is this. This argument does not argue why our time is special. The doomsday argument would have been valid since the dawn of the human species. The doomsday argument would also apply in the future.

Suppose we expect that we are halfway through the human population with 60 billion humans born and 60 billion left. Time passes, and another 30 billion people are born. We are at the 75% mark with 90 billion born and 30 billion left.

The 90 billionth human can use the same argument. I expect that I am in the middle of all humans who have or will exist. Therefore, I expect that we have gone through 50% of human population. 90 billion born, and 90 billion left.

Tick up another 30 billion. Are we at 120 billion gone, 0 left? 120 billion gone, 60 billion left? 120 billion gone, 120 billion left?

This seems contradictory. Unless there is some reason to privilege running the doomsday argument using now as time 0 instead of any other point in the past or future that I'm missing.

Or we could tick it backwards. The thousandth human would have correctly assigned a 1 in 600000000 chance of the human race making it this far. We have a 599999999/600000000 probability that didn't come to pass. It seems more likely the process for generating these probabilities is flawed.

Hide ↑Hide ↑

nadbor *says:*

June 4, 2019 at 8:55 am

That's not how it works. DA doesn't say you are exactly in the 50th percentile of all humans who ever lived. It says that a random human has 50% to be in the first half and

50% to be in the second half.

Or – to pick a different threshold – that the random human has 95% chance to be among the latter 95% of all humans and 5% chance to be among the first 5%. There is nothing in this that privileges the current time over any other time.

Assuming that only a finite number of humans will ever live and every one of them makes the prediction:

'I think I am among the last 95% of humans who will ever live'

– 95% of them will be correct.

This is a mathematical certainty. There is nothing metaphysical about it.

But it's unclear what we can do with this information. No one is a random human. Everyone is a specific human with additional information.

If everyone says 'I believe I am among the 95% tallest humans' – it is mathematically certain that 95% of the people will be correct and 5% will be wrong. But the little people who would be wrong to say that, know who they are!

If everyone somehow forgot how tall they are and were forced to make an estimate, saying 'I'm between 5th and 95th percentile of the distribution of human height' would be a sensible thing to do. But as long as you have any data whatsoever about your own life (your age, race, gender …), you can do better than that.

And so it is with doomsday.

Furslid *says:*

June 4, 2019 at 2:36 pm

I never said there was anything metaphysical about it. I do admit I simplified the math. However, the problem that there are wildly different estimates depending on which point we start counting from is still a problem. We are around the 60 billionth human life according to the wikipedia article. Someone is alive who is around the 50 billionth life. Between living humans now, there are equally good arguments for a 95% chance of 1 trillion humans max and a 95% chance of 1.2 trillion humans max. These cannot both be right, and this is a problem.

The fact that there are people who are wildly confident in extremely low probabilities is a problem. The thousandth person believes with 99.999% certainty that they are in the 99.999% last humans to live.

As to a random human saying "I am within the last 95% of humans to ever live. If every human who ever lived made that statement, 95% of them would be correct viewed from the end of time. However as the timeline progressed, the % of humans who were correct in making that statement would vary. When less than 5% of humans had been born, 0% were correct. Only when the last human has been born will the statement have been correct 95% or the time.

Hide ↑Hide ↑

nadbor *says:*

June 4, 2019 at 5:18 pm

> *Between living humans now, there are equally good arguments for a 95% chance of 1 trillion humans max and a 95% chance of 1.2 trillion humans max. These cannot both be right, and*

> *this is a problem.*

I don't see how this is a problem at all. The estimate of 'how many humans will live after me' is only as precise as the definition of 'will live after me'.You can specify that it means 'born after I was born' or 'die after I die' and you'll remove this ambiguity.

Yes, you'll get 1 trillion in one case and 1.2 trillion in the other, which if you assume constant human population size of 10bn translates into a difference of 20 years between the implied timeframe for the last human being born and last human dying. Nothing crazy about that.

> *The fact that there are people who are wildly confident in extremely low probabilities is a problem.*

Not really. That's how confidence intervals are supposed to work. 99.999% confidence interval should cover the true value 99.999% of the time and be off 0.001% of the time. If it covered the true value 100% of the time – it wouldn't be the 99.999% confidence interval!

There are 100k participants in a lottery and 1 prize. I'm 99.999% confident I'm not going to win and so should be everyone else playing. And yet someone wins every time. The winner is wildly confident in something (that they won't win) that then is proven wrong. Doesn't mean that the initial estimation of odds was wrong. Your beef is with statistics, not with DA.

> *Only when the last human has been born will the statement have been correct 95% or the time.*

Agree. This in no way invalidates it as a confidence interval

but it suggests that you can do better. This is the difference with the lottery example – participants in the lottery are assumed to be all the same and have no additional information. But if I rigged the lottery, I actually know for a fact I'm going to win. the 99.999% confidence in losing is still valid for a random participant but I'm no longer random.

Similarly, if I we detect a giant asteroid hurtling towards Earth about to hit it in 1 min, we would know for a fact that we're among the last 1% of all humans. At this point we're not randomly selected humans from the pool of all humans across space and time – we're the ones with a giant asteroid in front of us!

We don't exactly see a giant asteroid but still, we're not random samples, we can probably do better than to default to the ignorance prior.

Hide ↑Hide ↑

Furslid *says:*

June 4, 2019 at 7:48 pm

Yes, you'll get 1 trillion in one case and 1.2 trillion in the other, which if you assume constant human population size of 10bn translates into a difference of 20 years between the implied timeframe for the last human being born and last human dying. Nothing crazy about that.

There is absolutely something crazy about that. The estimates should be the same for everyone using the same methodology and data. Neither person has facts that the other doesn't, as both accept that the other was born when they were born.

Hide ↑Hide ↑

nadbor *says:*

June 5, 2019 at 1:57 am

You picked a weird thing to be concerned about. One estimate says that the last human will be born before the year 1trn at 95% confidence. The other says that the last person will die by the year 1trn + 20 at 95% confidence. This is perfectly consistent.

For starters, it is very easy to imagine a scenario where the last human is born on a space station in the year 1trn then runs out of supplies and dies 20 years later.

More importantly, these are confidence intervals, not clairvoyance. Of course you can construct different confidence intervals for the same quantity! Every scientific study reporting confidence intervals is open to the same objection – 'but you could have used a different procedure and gotten a different confidence interval'. Yes they could have. But they stick to a specific scheme to avoid p-hacking.

I have a 100-sided die. A valid 95% confidence interval on the outcome of a toss is (6, 100). So is (3, 97) so is (-1234, 95). No contradiction there. Each of those will turn out to have the true outcome in it 95% of the time.

Hide ↑Hide ↑

nadbor *says:*

All that said, I think you're on to something with pointing to the ambiguity with DA. I would like to explore a different example though.

Instead of using all humans as a reference class, I choose to use all humans born since 1987. I'm among the earliest born humans in this class. Let's say I'm the first. DA says that with 99% confidence there will be at most 99 more people born after me. But I know for a fact that there

already were billions!

Is the confidence interval wrong? No it's doing it's job perfectly – the same procedure gives the right result 99% of the time. But I have additional information and know that I'm in the 1% when the procedure fails.

You said that the problem with DA is that it treats the current time as special. I say the problem is that it doesn't. This time is special and so is every other time and we can do better than assume we're randomly selected from the pool of all humans.

Hide ⬆Hide ⬆

Furslid *says:*

June 5, 2019 at 10:30 am

Nadbor, that is another reasonable objection.

I actually need to withdraw some of my math. Because people were born after them, the 50 billionth person born knows that they are not in the last 1/6th of humanity. This would adjust their probability of being in the last 95% to 94%. So they are only 94% sure that the max human population is 1 trillion. The 60 billionth person is also 94% sure that the max human population is 1 trillion (they are 94% sure that they are in the last 94%.) Their 95% probable max population has increased to 1.2 trillion.

This also has a weird effect. Each new human born pushes off our projected end of the human race off by more than one person (if we are taking the chances of being in the first 95%, it increases the 95% confident max population by 20). So the doomsday argument has doomsday constantly getting further away.

This changes my view on the doomsday argument to

"True, but useless." It can never convince me that we are approaching doomsday in any reasonable timeframe.

Hide ↑Hide ↑

nadbor *says:*

June 5, 2019 at 1:06 pm

I agree with this.

Hide ↑Hide ↑

sty_silver *says:*

June 4, 2019 at 3:19 am

As someone who claims to understand the probability behind Doomsday, I think your intuition that it's nonsense is correct. I'm unsure whether your counter argument hits the root of the problem; that's something I'd probably have to think about quite a bit.

I've said in another comment that I think Bostrom's argument is solid. Does that one sound stronger to you? Not necessarily convincing, just more convincing than the Doomsday argument? If yes, I think that's pretty good news.

Hide ↑Hide ↑

kaminiwa *says:*

June 4, 2019 at 1:23 am

> *(This is the correct Bayesian action: if I know that a false argument sounds just as convincing as a true argument, argument convincingness provides no evidence either way. I should ignore it and stick with my prior.)*

I think there's two types of convincing: there's whether the argument is internally logically sound, and whether it can survive actual reality. I'm sure someone could construct an argument for cold fusion that was internally logically sound, but it's going to be extremely difficult to explain how they (or someone else) hasn't used this knowledge to corner the global energy market. If you want to push against an expert consensus, you first have to justify why the experts haven't changed their minds (in some fields, you can do this by saying "well, it hasn't been 30 years so the older generation hasn't died off yet", but quite a few people seem capable of actually-changing-their-minds)

If all arguments sound equally convincing-as-in-passes-the-"reality test" meaning something

has gone critically wrong. And if you think a few logically-consistent words can "prove" much of anything about reality, something different has gone wrong: gravity is still a theory, and we've done a fair amount of work to patch it's flaws in my lifetime alone.

(Tangentially, this is a lot of why I prefer to think in terms of "useful models" instead of "truth": Newtonian Gravity was wrong, but it was still a useful-enough model that we were able to land a man on the moon. Conversely, maybe the Earth really is flat, but knowing that doesn't seem to offer any actual useful value)

HideHide

kaminiwa *says:*

June 4, 2019 at 1:29 am

Pascal's Wager seems trivially countered by pointing out that there's an equal amount of evidence of the opposite hypothesis.

Or, in a more refined way: There is an effectively infinite number of theories that have equal evidence (namely: none), and consequently you should give each one only (1/infinity) attention, i.e. epsilon, i.e. basically zero.

HideHide

sty_silver *says:*

June 4, 2019 at 2:48 am

Two problems with this. A) it ignores priors, and B) it is simply not the case that there is zero evidence for the Christian god.

To elaborate on A): Yes, you can formulate countably infinitely many scenarios where not believing in god results in eternal bliss and believing in eternal suffering. However, they are increasingly complex and thus should have increasingly low priors. If you add up the probability of all of them, you should end up with something that converges, and probably to a rather low number. As a rough approximation, you could simply estimate that any other wild theories combined have some particular low probability, and all wild theories in the other direction have roughly the same, and then you're back at not having a response to PW.

To elaborate on B): there are people who, for example, claim to have entered heaven

during near-death experiences, and I even know an instance where they are said to have known details about heaven that are written in the bible but which they couldn't have known themselves.

Now, given how many people believe in Christianity, is it plausible that all of these are self-deceptions and perhaps lies? Absolutely. But are these instances some Bayesian evidence? Definitely. I mean, do you actually honestly assign the same probability to the flying spaghetti monster existing than you do to the Christian hell existing? I don't. They're both low numbers, but they're not exactly identical.

I recommend this podcast episode on the topic. It addresses your counter argument in a less technical way.

Hide ↑Hide ↑

LadyJane *says:*

> *I mean, do you actually honestly assign the same probability to the flying spaghetti monster existing than you do to the Christian hell existing? I don't.*

If by the Christian God you mean a literal old man with white robes and a long white beard who lives in a magical cloud palace, and by the Christian Hell you mean a literal fiery cave under the ground where living dead people are tortured by horned monsters with pitchforks, then yes, those ideas are about as realistic as the Flying Spaghetti Monster to me.

If your conception of God is some abstract Prime Mover responsible for the existence of the universe in some incomprehensible way, and your conception of Hell is some vague notion of karmic suffering that awaits the consciousness of sapient beings after death, then yes, those are *much* less ridiculous and unbelievable than the Flying Spaghetti Monster. Although the existence of Hell still seems incredibly unlikely, simply by virtue of the fact that we know human consciousness to be a product of human brains and there's no real evidence that it can exist independently of that material substrate.

Hide ↑Hide ↑

Randy M *says:*

June 4, 2019 at 7:48 am

> *If by the Christian God you mean a literal old man*

Why would he mean that?

Hide ↑Hide ↑

LadyJane *says:*

June 4, 2019 at 12:39 pm

Again, I was using "you" in the general sense. I'm aware that sty_silver's idea of God is probably not "old man in the sky," because theological arguments on this board tend to be rather sophisticated. My point was more that certain popular conceptions of God are every bit as ridiculous as the Flying Spaghetti Monster, or at least close to it. Russel's Teapot is not a particularly good argument against the deistic conception of God as some unseen and unknowable force outside of the universe proper, but it's a great argument against God as a Flying Spaghetti Monster or a white-robed sky wizard or any other discrete being with observable physical attributes located at some point in observable physical space.

Or, to state my argument as a more formal hypothesis: *The likelihood of a particular theory of God being true decreases exponentially with the amount of detail that theory has.* Maybe the Muslims and Ancient Hebrews were onto something with that whole "no graven images" rule!

Hide ↑Hide ↑

Randy M *says:*

June 4, 2019 at 12:44 pm

Literally nobody believes God is literally a bearded man in outer space.

Hide ↑Hide ↑

Deiseach *says:*

June 4, 2019 at 8:52 am

we know human consciousness to be a product of human brains

So we definitely know there is such a thing as consciousness, do we?

And that's not even getting into speaking of the soul, not the consciousness as such, which is the element that exists and suffers (until the General Resurrection when the body and soul are reunited for their eternal fates).

Hide ↑Hide ↑

LadyJane *says:*

June 4, 2019 at 12:58 pm

Alright, let's approach this from the opposite direction. Given the types of experiences that we know to be a product of the central nervous system, what are you left with if you take that system away?

Sensory awareness, including perceptions of physical pleasure and pain? Absolutely not, that's purely a result of our sense organs.

Memory? Again, definitely not, that's all stored in our neurological hardware.

Dreams? They exist as a result of sensory experiences and memories, so if you get rid of those, you get rid of the capability to dream too.

Emotions? This one's a little hazier, but still, probably not. Emotional states are largely tied to the release of various neurotransmitters and hormones in specific areas of the brain.

Thoughts? I suppose there's a little wiggle room here, depending on what exactly you count as thinking, but proper cognition (i.e. the ability to assess circumstances, derive conclusions from evidence and/or reason, make judgments and predictions, do logical and mathematical calculations, and so forth) is definitely a result of neurological processes. So once again, probably not.

Maybe there's still some pure singularity of consciousness awareness that remains after you strip everything else away, which could be described as a soul. But it's hard for me to see how it could be tortured in Hell: It can't feel physical pain, and it can't experience feelings like guilt or regret or sorrow or despair because it lacks memories and emotions and thoughts. Of course, you could just say that God creates a new body for the soul with all of the original body's memories preserved, but at that point you're making a lot of assumptions with no real evidence to back them.

Hide ↑Hide ↑

Deiseach *says:*

June 4, 2019 at 2:57 pm

Your argument works equally well against transhumanism and indeed cryonics: sorry guys, once the brain is mush there's nothing to be uploaded or thawed out or read or recreated, once the electrical energy stops flowing through the neurons that's it.

Now, myself personally, I *do* believe transhumanism and what-not are losing bets, but there are people out there (and indeed on here) who will argue that with you.

If you define "soul" to mean nothing more than "animating energy" then yeah, your definition holds water for your argument. But equally other people are arguing that all

this sensory awareness and the rest of it does not sum up to consciousness which is an illusory state and doesn't really exist.

So if what most of us are experiencing right now is basically a lie, then your "no such thing as a soul" argument holds just as true for "no such thing as consciousness" and it doesn't matter that we imagine we are conscious.

I certainly *feel* like there's an "I" here to experience things, but I could be wrong about that. I could be wrong about souls too, but we'll find that one out in the end when we die. Until then, I'm going to go on thinking I'm conscious right now, that an "I" exists, and that I have an immortal soul.

Hide ↑Hide ↑

LadyJane *says:*

June 4, 2019 at 3:49 pm

I believe that there's an "I" here to experience things too. It's debatable whether that "I" is purely a product of my current neurological state or something more, but it seems reductionist to claim that it's just an illusion and doesn't really exist at all.

And I'll admit, a broad claim like "consciousness is a fundamental *thing* in its own right" is very difficult to argue against. You can credibly argue that there's no proof it's true, but it's much harder to credibly argue that there's proof it's *not* true. At most, you can argue that specific traits that we associate with consciousness (awareness, memory, emotion, cognition) are known to be products of neurology, but it could still be argued that those things are merely the *content* of our conscious experiences and are distinct from consciousness itself.

So, regarding the hard problem of consciousness, I don't know the answer and I don't claim to know the answer. Indeed, I'm not even sure that it's the kind of question that could *be* answered, practically or conceptually. Much like asking "why does anything exist at all?," any possible answer seemingly leads to an infinite regress of more questions. That's part of why I'm an agnostic and not an atheist.

That said, once you start making specific predictions about the nature of consciousness and what happens to our "souls" after death, your position becomes much harder to rationally defend. Highly specific positive claims like that can be quite easily dismissed if there's a total lack of evidence for them.

Hide ↑Hide ↑
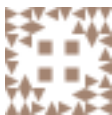
skaladom *says:*

June 4, 2019 at 5:23 am

Others have already remarked about how the very specific Christian worldview is but a point in a vast sea of imaginable worldviews, including ones compatible with survival of consciousness after death.

Ignoring the obvious problems of people who (by definition) did not die reporting about what death is like, here again we find ourselves with a clear case of overspecification: trying to infer a very specific and detailed chunk of information (the whole of Christianity) from a relatively weak and generic hypothesis (survival of consciousness).

If we were to take the basic hypothesis (testimonies of survival of death) seriously, the obviously minimal consequence one would draw would be a kind of idealistic position, where the material world is a collective dream by a set of ultimately disembodied consciousnesses. Variants of this have been elaborated in ancient Greece, ancient India, and "classical" German Idealism. Note that this is compatible with conscious survival at death, but does not necessarily imply it: the individualized stream of consciousness could just as

well die together with its dream-body. I actually give this general kind of view a good 50%.

To go back to the other point, the cool thing about overspecification is that it happens pretty much everywhere, because it's a generic dynamic in a cultural species such as ourselves.

Imagine our prototypical proto-humans before the invention of agriculture. Obviously it pays to know which bark can be pounded and taken as a painkiller, and also which ones need to be avoided at all costs because they are poisonous. So soon enough every tribe has figured out the use of a whole bunch of their local medicinal plants.

But then someone comes and randomly claims (maybe as a joke) that this specific bark works better if taken while crouching, and by the way there should definitely be no cows present. This does not actually help with the purpose in any way, but it doesn't really deter from it – the conditions are easy enough to fulfill. More surprisingly, this new piece of "information" has remarkable sticking power: the person being given the advice feels like they got an extra bit of thoughtful care; the one giving it gets some extra prestige for knowing such an obscure fact, and soon enough the whole tribe has yet another reason to rationalize that the neighboring tribe, who does not crouch or avoid the presence of cows when taking their medicine, are uncouth savages utterly unlike us, and therefore absolutely deserve to be looted and treated like animals.

Multiply by a few thousand years and you get… well, human culture!

Hide ↑Hide ↑

DragonMilk *says:*

June 4, 2019 at 10:08 am

So contrary to popular belief, the Christian claim is actually bodily resurrection, in contrast to traditional Greek thought that the soul escapes the impurities of the body. The survival of consciousness bit is Greco-Roman, not Christian. The Christian claim is that like Jesus, resurrection will be bodily.

As you can imagine, that claim sounds as absurd then as now.
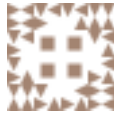
skaladom *says:*

June 5, 2019 at 11:27 am

Thanks for the clarification. I was just going with the flow of the previous poster. If bodily resurrection rather than conscious survival is indeed what Christianity really claims, then I really don't see how luminous near-death experiences can be adduced as proof...

Then again, it's nothing too unusual for views that are not the orthodox institutional position to actually hold more sway than those that are, I guess.

kaminiwa *says:*

June 4, 2019 at 8:56 pm

Re: Priors: "God" and "Hell" are two of the most massively complex ideas ever to be entertained by humans. An appeal to simplicity is an argument *against* God existing.

"These two cancel out" was my point: if you accept the premise "God's truth is unknowable", which is an explicit part of Pascal's Wager, then one should remain staunchly neutral and neither believe or disbelieve. Whereas Pascal's Wager says that in the absence of a knowable truth, you should believe, which is nonsensical given the premise.

—

Re: Evidence: If one wishes to toss out the "unknowable" premise of Pascal's Wager, then one has to weigh "people believe" against all the evidence against any specific god. And that scale is drastically weighted against religion.

I'd also argue that, looking at history, "a large group of people believe in this" has a track record so bad that I'm not sure it actually *is* Bayesian evidence on any meaningful scale.

> *I mean, do you actually honestly assign the same probability to the flying spaghetti monster existing than you do to the Christian hell existing?*

Mu – you're asking the wrong question:

If we assume there is a divine, I can make all sorts of arguments for why Christianity is more likely than the Flying Spaghetti monster.

But the probability of there being a divine is so microscopic, that we're arguing over grains of sand on the beach.

So, there is a difference, but they're still both so radically unlikely that there's no real point to thinking about them (except as a thought exercise or reference point)

Hide ↑Hide ↑

Faza (TCM) *says:*

June 4, 2019 at 1:40 am

> *Even the smartest people I know have a commendable tendency not to take certain ideas seriously. Bostrom's simulation argument, the anthropic doomsday argument, Pascal's Mugging – I've never heard anyone give a coherent argument against any of these, but I've also never met anyone who fully accepts them and lives life according to their implications.*

It appears to be an immutable law of the universe that if someone writes a post mentioning, say, FizzBuzz on a programming blog, the comments are going to be full of people's solutions to FizzBuzz.

Similarly, I expect you *knew* that if you put something like the above in an SSC post, you're going to get a bunch of counter-arguments to the above.

That out of the way, here's mine:

Pascal's Mugging

I can't help but think of PM as what happens when a bright mind discovers a new tool and begins to run with it. Unfortunately, I don't think EY stopped to consider that the math doesn't work in his favour when writing [the post](#).

At the heart of the argument is the calculation $EV = V_n * P$, where EV is the expected value, V is the value of a single life, n is the number of lives snuffed out and P is the probability of this happening. EY is arguing that if n is some huge number like 3^^^^3 (Knuth's arrow notation is the shiny new tool in this case), there's no way we can make P a small enough number to bring EV to less than the $5 demanded.

Hold my beer and watch this!

The first approach requires the acceptance of the following premises:
1. In order to kill 3^^^^3 people, the mugger must kill person 1 *and* person 2 *and* person 3, etc., etc.,
2. The mugger can kill people in any order and any number of steps, so killing person 2 doesn't depend on killing person 1, killing person 3 doesn't depend on killing person 2 and/or person 1, etc., etc.

I hope neither is particularly controversial.

Having accepted these two premises, we can approach assesing the probability that n people will be killed as the probability of n (by 1) independent (by 2) events, where each event is the killing of one person.

For simplicity, we might assume that we consider it equally probable that person i will be killed as person j being killed – in other words: we do not diminish our assumed probability of the next person being killed. What probability can we assign to n people being killed?

Let probability $P(i)$ of each person i being killed be $1/p$, where $p > 1$. The probability of n individual people (independent events) being killed is $1/p^n$.

Oops…

Let's put this back into the expected value calculation we're working from and plug EY's Big Number in it for good measure:

$$EV = (V * 3\text{^^^^3}) / p\text{^}3\text{^^^^3}$$

Two things jump out here:

1. For any $p > 1$, $p\text{^}3\text{^^^^3}$ is going to be vastly greater than 3^^^^3,

2. The more arrows we jam in there, the lower our *EV*, because *V* is constant.

You're better off credibly threatening just one person because $1/p$ is gonna be greater than $1/p\text{^}n$ for any $n > 1$.

But maybe you're not convinced that treating killing *n* people is equivalent to independently killing idividual people *n* times.

Here's approach 2:

What if our mugger fails to kill 3^^^^3 people, but only succeeds in killing $3\text{^^^^3} - 1$ people? How about $3\text{^^^^3} - 2$? How about $3\text{^^^^3} - 3\text{^^}3$?

There may be no kill like overkill, but clearly 3^^^^3 people is way beyond the kind of threat that would make most folks part with $5 – if it is credible. Similarly, I think few people would seriously argue that "since you failed to kill exactly 3^^^^3 people, I was right in keeping my $5". Let's consider *this* angle.

If the mugger kills only 1 person, they haven't killed 2 people, or 3, or 3^^^^3. Similarly, if they have killed only 5 people, they haven't killed 15, etc., etc.

We can therefore view the killing of *m* people, where $m <= n$ as one of a series of mutually exclusive events.

We can assign probability *P(i)* to each such possibility, so *P(1)* is the probability that 1 person is killed, *P(2)* is the probability that 2 people are killed and *P(3\text{^^^^3})* is the probability that 3^^^^3 people are killed.

How do we assign probabilities across the set? Actually, we don't have to.

For most people, the value of a single human life greatly exceeds $5. For this reason, a rational mugger needs only to credibly threaten *one* person, in order to extract the ransom. If threatening one person isn't enough, then maybe threatening two will be. Or maybe three. It's reasonable to expect that the actual number lies somewhere between 1 and n, because if threatening n people won't do the job – nothing will (in our set of possibilities).

Let $P(i) = 1/pi$, where $P(1) = 1/p1$ and $pi >= p1$. We're talking n mutually exclusive events (one for each possible number of people killed), so P = sum of $P(i)$ over n, with the best case (greatest probability) being $P = P(1) * n$ (in other words, we consider it equally likely that the mugger will kill 1 person as 3^^^^3 people):

1. $EV = Vn * P(1)n$
2. $EV = Vn / (p1 * n)$
3. $EV = V / p1$
4. $EV = V * P(1)$

I've gone the roundabout way to hammer the point in. Funnily enough, it seems like less is more if you plan to threaten people for concessions.

I'll leave it up to you to decide what that tells us about the world and ourselves.

HideHide

Said Achmiz *says:*

June 4, 2019 at 2:16 am

> *For simplicity, we might assume that we consider it equally probable that person i will be killed as person j being killed – in other words: we do not diminish our assumed probability of the next person being killed.*

In fact we should *increase* our assumed probability of the next person being killed. The more people we know the mugger has killed, the more likely it is that he's used some sort of mass killing device, with which it is much easier to kill the marginal person. (There are also other considerations—psychological, sociopolitical, technological, cosmological, etc.—all of which make each marginal person [across many, though not necessarily all, regions of the number line up to our target number] more likely to have been killed.)

This tanks your calculation.

Faza (TCM) *says:*

When the killing has already started, it's a bit too late, don't you think?

Pascal's Mugging has us evaluating a postulated, but unproven, capability of "us[ing] [...] magic powers from outside the Matrix to run a Turing machine that simulates and kills [n] people".

Leaving aside that most people are likely to collapse the probability of using "magic powers from outside the Matrix" to essentially 0, I am interested to know why we should consider the claim "I will use a Turing machine to simulate and kill one person" to be less likely than the claim "I will use a Turing machine to simulate and kill ten people", when all we have to go on is that person's word.

I mean, just going off of what we know about Turing machines, simulating ten people requires more resources than simulating one.

You're assuming the mugger has already demonstrated killing people. You made that up, I'm afraid. The scenario explicitly assumes we don't *know* if the mugger can kill 3^^^^3 people – or even anyone at all.

Said Achmiz *says:*

No, you misunderstood me entirely.

You are saying that if the probability of being able to kill 1 person is p, then the probability of being able to kill 2 people is p^2. (And so on, for arbitrary $n > 2$.)

And I am saying that is false. That's all.

I do not assume the mugger has demonstrated anything, or that

any killing has already taken place.

Simply, if we *suppose* that the mugger *can* kill 100 people—well, whatever probability we assign to *that*, deriving the probability that he can kill 101 people should *not* involve multiplying that kill-100-people number by the probability that he can kill just *one* person! That is far, *far* too conservative—for the reasons I listed.

Hide ↑Hide ↑

Faza (TCM) *says:*

June 4, 2019 at 5:01 am

Let's see the reasons you listed:

> *The more people we know the mugger has killed, the more likely it is that he's used some sort of mass killing device, with which it is much easier to kill the marginal person.*

> *(There are also other considerations— psychological, sociopolitical, technological, cosmological, etc. —all of which make each marginal person [across many, though not necessarily all, regions of the number line up to our target number] more likely to have been killed.)*

(Edited for clarity.)

You can't simply assert such things and expect them to be taken at face value. Let's try the easy one:

> *The more people we know the mugger has killed, the more likely it is that he's used some sort of mass killing device*

That most certainly does not follow from the scenario as stated. The threat was:

> *I'll use my magic powers from outside the Matrix to run a Turing machine that simulates and kills [...]*

There is absolutely nothing about a "Turing machine that simulates and kills" to suggest that the probability of it being used to simulate and kill n+1 people is greater than that of it being used to simulate and kill n people. It's not – to touch on niohiki's objections below – a bomb. For all we know – and this is consistent with our knowledge of computers – the machine works by spinning up a "person" process and subsequently terminating it, possibly with a (subjectively) gruesome shutdown sequence. Moreover, as it is a [Turing machine](#), sequential running of the simulated programs is just as (if not more likely) as parallel execution (that would imply Turing *machines*).

A basic, no-frills reading of the scenario is "I will run the 'simulate-and-kill' program on my Turing machine 3^^^^3 times".

As for the rest of your reasons, you'll need to unpack them first.

Hide ↑ Hide ↑

Said Achmiz *says:*

June 4, 2019 at 5:26 am

> *... that the probability of it being used to simulate and kill n+1 people is greater than that of it being used to simulate and kill n people.*

This is not what I said. I understand the temptation toward

carelessness, but the subject demands precision.

As for the rest… your objections fall well within the bounds of Dennett's famous quip about mistaking a failure of imagination for an insight into necessity.

Perhaps more importantly, you have (as do most commentators on this subject) quite misunderstood the purpose of the original Pascal's Mugging thought experiment. To critique the scenario for purported *particular* incoherence is to miss the point.

Faza (TCM) *says:*

June 4, 2019 at 7:16 am

> *Perhaps more importantly, you have (as do most commentators on this subject) quite misunderstood the purpose of the original Pascal's Mugging thought experiment. To critique the scenario for purported particular incoherence is to miss the point.*

The purpose of the original Pascal's Mugging experiment was, to quote EY:

> *But suppose I built an AI which worked by some bounded analogue of Solomonoff induction – an AI sufficiently Bayesian to insist on calculating complexities and assessing probabilities, rather than just waving them off as "large" or "small".*

> *If the probabilities of various scenarios considered did not exactly cancel out, the AI's action in the case of Pascal's Mugging would be overwhelmingly dominated by whatever tiny differentials existed in the various tiny*

It's not my fault he made a hash of it, because he doesn't understand AI or the matters he chooses to expound upon.

I have given two plausible approaches to making the "probabilities cancel out". Nothing therein requires anything that is not contained in the original scenario. An AI integrating this approach will not fall for Pascal's Mugging and this requires nothing but "calculating complexities and assessing probabilities".

Unless you're seriously worried about magic powers from outside the Matrix, your objections are spurious.

The solution is simple and I have given you a concise explanation, supported every step of the way. You have chosen to instead postulate issues that are not part of the scenario and accuse me of misunderstanding – also by pure assertion – and I would honestly be more surprised if I didn't know this is your standard MO when the Sequences are involved.

This conversation has ceased being productive. Good day, sir.

Hide ↑Hide ↑

niohiki *says:*

June 4, 2019 at 2:30 am

Erm, I also don't get the balance of probabilities in the argument for PM. Mostly because there are hyperpriors to be considered (like what is mentioned in the post: how sure can I be that these probabilities make sense, that I am not being fooled, that I have understood all the consequences, that my action will not in fact have exactly the opposite effect through some unlikely, yet *similarly* unlikely, mechanism, etc etc).

But... the probabilities here are not independent and multiplicative. If there is a probability $p=0.5$ that I push a Big Red Button that drops nuclear weapons over all of Europe, then I shouldn't argue that it's fine because there's a probability $0.5^{700.000.000}$ of killing all those people, or whichever sum 0.5^m over m my bounds are happy with. There is about $p=0.5$ that everything becomes terrible, and that's it. And I really shouldn't push the button. Of course, you can *define* some q per person such that $q^{700.000.000} = p$. But then one is saying "q is arbitrarily close to 1" and not thinking about how pathological "arbitrarily close to" is in the same way that the PM argument makes a magic step and does not consider properly limits when placing the "arbitrarily large" payoff vs a "arbitrarily small" probability. Also, this q doesn't make any sense because it does not represent any actual probability. Really, most of the time random variables *are not* independent.

Hide ↑Hide ↑

Faza (TCM) *says:*

The Pascal's Mugging scenario assumes we *don't know* what the probability is.

EY postulates that the sheer size of the downside overrides whatever (im)probabilities we may assign.

I'm saying that not only is this not the case, but the (im)probability scales with the magnitude of the claim. If I meet a mugger on the street, it's much more probable that they will be able to kill one person (all they need to do is pull out a knife and stab the nearest passer-by), than that they will be able to kill everyone in the city, which is more probable than that they will be able to kill everyone in the country, which in turn is more probable than that they will be able to kill everyone in the world.

We have no good way of estimating the difference in probabilities – especially when it comes to stuff like "magic powers from outside the Matrix". EY was looking for a mechanical fix to the problem and this is it.

Approach 1 works under the assumption that we assign some non-zero probability to a person being killed and scale along the number of people the mugger is claiming they will kill.

Approach 2 considers the possibility that some number of people will be killed and works off that.

In both cases, note the result is the same: if we consider it remotely plausible the mugger will kill just one person, the mugging succeeds, because $5 is insignificant compared to the value of a human life.

That's why muggers carry knives, as opposed to magic powers from outside the Matrix.

Hide ↑Hide ↑

niohiki *says:*

June 4, 2019 at 3:38 am

> *The Pascal's Mugging scenario assumes we don't know what the probability is.*

I know, I know, that's what I wanted to point out with the magical step. My problem with PM is what I and others in this thread and elsewhere have said: one cannot say "this infinity is always more infinity than that infinite-ly small zero". You really need a model for how zero the zero is and it includes things like "the PM argument actually makes no sense but I could still be convinced by it because I'm not a perfect reasoning machine", which have *pretty high probabilities*.

But stating that

> *but the (im)probability scales with the magnitude of the claim*

and using a mugger as an example… does it really follow? Going back to nuclear weapons: the unlikely claim is "the nuke will kill about 5 people", instead of "the nuke will kill thousands". We could also be talking about someone in a position of power implementing policy ("this bill on free speech will ruins the lives of millions of

people" vs "this bill on free speech will ruin the lives of some / people") and so on.

To use the slang,

> but the (im)probability scales with the magnitude of the claim

is proving too much, as you use it. Sure it works for a *mugger*. But if anything "big" automatically has low probabilities, then there would be no declarations of war, because if we model a war like a mugger that kills one person after another, it is *so unlikely* to end up with Stalingrad that it clearly must have never happened, and it is more likely that everything we know about it comes from some Communist/Bilderberg/Neocon conspiracy for some unlikely reasons, but not as unlikely as a mugger killing a million people in a row.

Hide ↑Hide ↑

Faza (TCM) *says:*

June 4, 2019 at 3:55 am

I see where you're coming from and I don't disagree, in principle.

I would like to point out to you, however, that there is a lot of knowledge you are tacitly introducing into your examples. Knowledge that we simply don't have in Pascal's Mugging.

Take nuclear weapons. Everyone who mattered when nukes were introduced had a pretty good idea of how they worked, what they were capable of and was prolly working on their own. The Japanese certainly were.

The evaluation of any claims made about nuclear weapon capabilities is necessarily – and rightly – coloured by what we already know about nuclear weapons.

The same goes for modeling a war. We know how wars work and therefore our estimates of casualties and damage will be coloured by what we know of our own and the enemy's capabilities.

In both cases, we don't do simple models primarily because we have better models.

In Pascal's Mugging we have absolutely no knowledge, so it's simple models or pulling numbers out of a hat. I prefer simple models, 'coz the rabbits ate all my numbers. Scaling the improbability with the size of the claim – in the absence of any prior knowledge – is consistent with what we know about how the world works (more work requires more energy).

That isn't the *point* though, the point is that Pascal's Mugging is a misdirection, a sleight-of-hand (sorry, Robert Jones). Pascal's Mugging is pointing at the Big Number and waving agitatedly, whilst ignoring the fact that the same effect can be achieved with a much smaller number.

Hide ↑Hide ↑

niohiki *says:*

> *Knowledge that we simply don't have in Pascal's Mugging.*

True… but neither do we have knowledge to suppose that probabilities are, indeed, independent, so I should not really have the right to reason based on things like "the multiple events are independent". Looking back I probably was not very clear, but I was not affirming that the Horrible Thing With Which We Are Blackmailed would happen as a nuclear bomb drop, or denying that it would happen as a lot of independent killings. I just wanted to

provide an example of plausible situations where probabilities are indeed not independent, and *precisely* because we have no idea, we cannot assume that they are for the analysis.

But we don't need that analysis! *Precisely* because we have no idea we cannot either just say that yeah, sure, there exists a sector in possibility space for which outcome utility grow faster than the corresponding probability decreases. Why?

> *Pascal's Mugging is pointing at the Big Number and waving agitatedly*

Exactly! That's all the problem we need to point out! Sure, it is easy to make utilities that seem arbitrarily fast-growing (Throw more people into the VR-hell! More!). But why, seriously why, do people assume with all tranquillity that probabilities *cannot* diminish equally fast, or even faster? Why can they see that numbers can get large very fast, but not small even faster?

Hide ↑Hide ↑

Faza (TCM) *says:*

June 4, 2019 at 7:18 am

So… I don't get it. It seems we are in violent agreement.

Hide ↑Hide ↑

niohiki *says:*

June 5, 2019 at 12:12 am

Well, mostly. Just not about that *particular* reasoning against Pascal's Mugging.

Hide ↑Hide ↑

Faza (TCM) *says:*

The Doomsday Argument

A deductive argument is sound if it is logically valid *and* all of its premises are true. For our purposes here, I will assume the Doomsday Argument is valid.

This leaves the issue of its premises being true. For any kind of prediction being made via the DA, the easy target is the number of humans who ever lived (as borne out by the linked Wikipedia article), but I find it more fruitful to attack the other one:

> *we could assume that we could be 95% certain that we would be within the last 95% of all the humans ever to be born*

The DA is timeless (in the sense that it is not dependent on when we ask the question) and the premise in question reflects this (by fudging the numbers). Therein lies the problem.

Because the DA is timeless, it can just as well be asked by people who *are* in the first 5% of all the humans being born. For them, this premise would be false, as would any conclusions drawn from the DA. In order to know whether this premise is true or not with respect to themselves, the asker would essentially need to know the answer to the question they're trying to answer via the DA (and thus have no use for it).

If it didn't take more work than I'm willing to put in, we could attempt to answer the question "at what point in time were the last people for whom the argument was demonstrably false alive". We can demonstrate the argument was false for those people through the power of hindsight – we *know* there were more people alive up till the present day than they could have predicted by applying the Doomsday Argument.

If the Doomsday Argument can be demonstrated to have given false answers in the past, there is no reason to expect it to give true answers for us today.

The problem with the "we're part of the last 95% of humanity" premise isn't that it's false for us. It is – to use a computing term – null. We cannot reliably assert its truth value either way. (Mis)applying the principle of null propagation, any expression – and therefore, also the Doomsday Argument – with a null term evaluates to null.

The Doomsday Argument is fun, but it holds no predictive value.

faul_sname *says:*

June 4, 2019 at 5:05 pm

Let's change around the terms of the doomsday argument a bit.

95% of employees at any given company will be among the last 95% of employees hired by that company. If you are hired by a company, then absent other information your prior should be a 95% chance that you are in the last 95% of employees that company will hire. Change out "be hired at a company" for "be born into the human species" and you have the classic doomsday argument.

That argument is pretty much a tautology, and is thus not arguably not that useful, but it's pretty obviously *true*.

There is at least one case where the analogy actually *is* a useful way of thinking about something: pyramid schemes. If only the first 10% of people to join the pyramid get rich quick, then if you decide to join the pyramid scheme, you should assume there is a 90% chance that you are among the 90% of people who will join last and thus there is a 90% chance that you will not be getting rich today.

The doomsday argument does not tell you how to update your estimate of how likely you are to be in the last n% of humans ever born – rather it tells you what your prior should be before taking any other evidence into account.
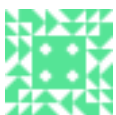
Robert Jones *says:*

June 4, 2019 at 3:04 am

> *The mugger can kill people in any order and any number of steps, so killing person 2 doesn't depend on killing person 1, killing person 3 doesn't depend on killing person 2 and/or person 1, etc., etc.*

> *we can approach assesing the probability that n people will be killed as the*

> *probability of n (by 1) independent (by 2) events, where each event is the*
> *killing of one person*

This is a sleight of hand. Killing person 2 may not *depend* on killing person 1 (in that Omega could hypothetically kill person 2 without killing person 1) but they're not probabilistically independent events.

Hide ↑Hide ↑

Faza (TCM) *says:*

How so?

Hide ↑Hide ↑

Ristridin *says:*

There's a distinction between causal independence and probabilistic independence. If A and B are events, then B may causally depend on A if event A directly causes event B (I don't have more precise terminology for this). Causal independence between A and B then means that A does not cause B, nor does B cause A. Causal independence can be extremely difficult to show in practical situations, but is clear if A and B can occur in any order. This is more or less what the first quoted statement asserts.

On the other hand, A and B are *probabilistically* independent if the probability that both A and B occur is equal to the product of their respective probabilities (or more concisely: P(A and B) = P(A) * P(B)). Probabilistic independence is a precise mathematical term, and is what you use in the second quoted statement.

However, causal independence is much, much weaker than probabilistical independence. That's the 'sleight of hand' mentioned; your argument uses these two meanings of independence as if they were interchangeable.

Example: Two people are in a room, and a bomb goes off with

probability 0.5, killing both if it does. Then event A (person 1 dies) is *causally* independent of event B (person 2 dies), since if the bomb goes off, A and B can happen in either order. However, A and B are not *probabilistically* independent; P(A) = 0.5 (there is a 0.5 chance of person 1 dying, namely the probability that the bomb goes off), P(B) = 0.5 (there is a 0.5 chance of person 2 dying), but we don't have P(A and B) = 0.5*0.5 = 0.25; we still have P(A and B) = 0.5. The bomb does not become less likely to go off if there are more people in the room.

In other words, the claim that the probability of the mugger killing n people is p^n where p is the probability of the mugger killing 1 person would require a lot more evidence than premises 1 and 2.

As to your approach 2, while I can't quite read your notation in the latter part (what are EV and Vn? Is EV expectation value, V the value of an individual, and Vn = V*n?), I think that at some point you claim that the best case scenario is that the probability P(n) of Pascal's mugger killing n people is equal to the probability P(1) of Pascal's mugger killing 1 person.

This, I do not quite agree with, as the worst case is that P(3^^^3) = p > 0 where p is the probability that the mugger kills at least one person (and all other probabilities are 0). Compare to the bomb example; the probability the bomb kills exactly two people is 0.5, while the probability it kills exactly one person is 0. I think your argument relies on an implicit assumption that the probability that Pascal's mugger kills exactly n+1 people is smaller than the probability the mugger kills exactly n people, which need not be the case.

Note that I do agree that in any practical situation, threatening to kill 1 person (specifically: the target of the mugging) is indeed usually sufficient to get $5, which is why most muggers are not Pascal's muggers. In fact, I would consider any claim of being able to kill 3^^^3 people so implausible to not be worthy of consideration (much less plausible than a threat to kill me

personally). This does not however counter the actual content of the argument that Pascal's mugging makes, which is that acting based purely on expectation values to maximize utility (rather than, say, use 'common sense' (however you would define that)) may cause one to act in a manner that we would consider stupid.

Hide ↑Hide ↑

skaladom *says:*

Nice mathematical treatment here. My math is hopelessly rusty, and I know this refutation thread is completely tangential to the main post, but here I go anyway. My point of attack is that the usual set of (supposedly) rational arguments with (supposedly) unsettling consequences is that thought is able to build huge castles in the air (including all sorts of asymptotic zeros divided by other asymptotic zeros), but at some point correspondence with experiential reality just gets completely lost. I don't think it's particularly hard to pinpoint examples.

For simulations, the hidden premise is that simulating an entire world (or astronomically worse, an entire universe or multiverse) is feasible within another world, and so on with no limit. But a calculation requires resources, and I see no particular reason why *fully* simulating a billion particles would require less than a billion particles in the first place. Maybe another hidden assumption is that maybe one can short-circuit some of the complexity of the host world when simulating a hosted world, in the same way that there is a whole area of research trying to make brain-inspired computation devices where neurons are replaced by simple logical gates or basic mathematical functions. Yet I see, once again, no reason to think that any of the layers of complexity of the world would be irrelevant to the end result, and of course the possibility of grossly incomplete simulated world brings none of the paradoxical consequences of the original idea, because it fails to create an infinite regress.

And without huge numbers of cheap world-simulations, there goes Pascal's mugging.

For the "doomsday" reasoning, others have very well put into doubt the very possibility of reasoning on the probability of "being born as me", which btw implies a conceptual duality (between something that is born, and a "me" that it becomes) which is completely absent in reality. So I'll just add that according to the quoted Wikipedia article, the argument suggests that humanity might be around for another

~9 millenia. While certainly much less than the usual optimistic takes on humanity, and less than the naive view that we've been around for hundreds of thousands of years and should at least last about as much, it is in any case such a large number that I cannot see how it would cause emotional concern in the present. Just think, did the ancient Sumerians have any inkling of what our present existential worries might be? We're talking of an even greater timespan, so if something might come and terminates humanity some 10,000 years from now, chances are our current cultures have not even remotely been able to imagine it.

Another flaw with the doomsday thing is that it assumes that there is a clear beginning to "us" as humanity, rather than a very gradual process of evolution. And in any case, wouldn't it make more sense to put the "us" at the origin of sentiency, rather than the specific homo sapiens species? If so, that would substantially raise the past number of beings, and accordingly raise the (supposedly) expected number of future sentient beings on earth…. it begins to look even like an optimistic outlook, given the current strength of global warming!

Hide ↑Hide ↑

skaladom *says:*

June 4, 2019 at 4:13 am

Even less on topic, I'll add that the same sort of common sense responses can easily be given to abstruse metaphysical arguments such as those that purport to prove the existence of "God". Basically, starting from very basic and nearly undisputable premises, you go on to prove something hopelessly generic and contentless, such as "the ultimate nature of the universe, whatever it may be, is at least as real as the universe itself"… and then proceed to project onto it the whole mythology of your favorite religion with god(s), afterlife, salvation and the whole lot.

And I find it even more disturbing to consider how our conceptual picture of something as familiar and basic as matter is hopelessly circular! We think we somewhat "understand" matter because our familiarity with popularized basic physics tells us that it is made of particles bound by forces. And quantum physics says forces amount to the exchange of particles, so it all comes down to particles. But how do we think we understand particles? Unless you're a hard-nosed theoretical physicist with a grounding in philosophy of science,

for whom a particle might be "that which answers to a certain equation", most of us just think of them as either hard solid balls (classical model), or as fuzzy cloud-like elastic things (quantum picture)... and in both cases the basis for our mental image is none other than ordinary matter, the very think we're trying to figure out to begin with! Try to imagine what matter actually may be, without recourse to circular imagery – your mind might well go blank.

Hide ↑Hide ↑

vaniver *says:*

June 5, 2019 at 10:48 am

> *2. The mugger can kill people in any order and any number of steps, so killing person 2 doesn't depend on killing person 1, killing person 3 doesn't depend on killing person 2 and/or person 1, etc., etc.*
>
> *I hope neither is particularly controversial.*

Suppose my plan is to kill 8 million people at once by destroying New York City with a particularly large explosion. If both person 1 and person 2 live in New York City, then I can't kill one and not the other using this method. They are not independent events.

[This is basically Said's objection, which is entirely valid, with a concrete example.]

Hide ↑Hide ↑

tmk *says:*

June 4, 2019 at 2:14 am

This is a classic and good, but I could not imagine you writing this today. In the 2019 SSC climate it's just not ok to be off-handedly positive about education.

HideHide
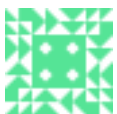
Robert Jones *says:*

June 4, 2019 at 2:58 am

> *A friend recently complained about how many people lack the basic skill of believing arguments. That is, if you have a valid argument for something, then you should accept the conclusion. Even if the conclusion is unpopular, or inconvenient, or you*

don't like it. He envisioned an art of rationality that would make people believe something after it had been proven to them.

I'm concerned that "if you have a valid argument for something" is equivocating between "you are convinced the argument is valid" and "the argument is objectively valid". I think it's probably very unusual for someone to be fully convinced of the validity of an argument (and the truth of the premises) but to reject the conclusion.

If the problem is really that people aren't convinced by arguments which seem to your friend to be valid, then epistemic modesty should require your friend to acknowledge the possibility that it is he who is mistaken.

There are lots of counterintuitive results in maths, but when people are shown the proofs (and assuming they can follow the proof) they're convinced. I would suggest that this is because following the proof enables one to see not only that the result is true but also *why* it is true.

HideHide

NoRandomWalk *says:*

June 4, 2019 at 8:08 am

> *Even if the conclusion is unpopular, or inconvenient, or you don't like it.*

is the context for the type of arguments we're talking about.

Just so you get the context (it's not math proofs) often these are arguments which have moral or philosophical implications. The rationalist idea you should 'take ideas seriously' became foundational to the culture in the context of situations such as someone who normally donates to her local soup kitchen with her time, being persuaded that if she really cared about the poor she would become a high powered laywer, and then choosing to both not do so and continue to tell herself that her primary sacred value is helping the poor.

I take your point, however

> when people are shown the proofs **(and assuming they can follow the proof)** *they're convinced*

is doing a lot of work. I have degrees in math, and there have been several times that

I genuinely thought I understood and followed a proof and yet the conclusion turned out to be wrong.

Hide ↑Hide ↑

Deiseach *says:*

June 4, 2019 at 3:04 pm

Eh. If everybody who would otherwise feed the poor instead becomes a high-powered lawyer, then who's feeding the poor right now? I guess if you die of starvation, the high-powered lawyers can take a case against the council or the government or somebody for not taking care of you, but that's not much consolation to the dead person.

We need the high-powered lawyers to take the cases to force the social services to provide meals, but we also need the people to cook and disburse those meals to the poor on the ground as well. "Too many chiefs and not enough Indians" is what comes to mind about that kind of rational "everybody become a high-powered lawyer" advice.

Hide ↑Hide ↑

sty_silver *says:*

June 4, 2019 at 3:04 am

I… disagree with you. What a strange feeling.

It seems to me that neither policy a) "you should believe arguments that sound convincing" nor policy b) "you should not believe arguments that sound convincing" are good policies. It simply depends on the domain.

For example, suppse I conduct a plausible sounding argument about why Tesla's stock price is guaranteed to hit at least 300$ at the end of the year. Lots and lots of very smart people are thinking about this and disagree with me, and moreover, they would absolutely act on their beliefs if they agreed with me. And I'd notice if they did, because I can just type "tesla stock" in google and see what most experts think. (This is the same example EY uses in Inadequate Equilibria as to when you should apply epistemic humility.) Therefore, you should not believe this clever argument.

Now consider Bostrom's simulation argument. Have lots and lots of smart people thought about

this? Maybe. What do most of them think? I have no idea. Would we know it if they genuinely believed it? No. Moreover, there are reasons to expect that this argument is going to get less support than it should, because it sounds very strange and many people will have a knee-jerk rection towards it — but, and this is important, there really isn't any rational reason for why this particular hypothesis should receive an exceptionally low prior (it maybe should have a prior < 50%, but not 50%.

So should you believe the argument, or at least assign it some significant probability? I think so. I also hereby claim that I fully accept the argument and live by its implications (though I don't think there really are any implications).

Let's take the Doomsday argument, because I happen to be pretty sure that it is inaccurate (I've read and written coherent(?) rebuttals). Much of the same as above applies: there's no super strong reason to assume it has to be false, you don't really know what the smart people believe, and you don't even really know how many smart people have thought about it. So if it does in fact soud convincing to you, should you assign it a significant probability? I think so. You'd arguably go wrong in this case, but well, a reasonable metric won't work every time.

In my case, even before being convinced by the rebuttals, the argument *hasn't* really sounded particularly convincing. But I can concede that for many people, there isn't a substantial difference between how they react to Doomsday an the simulation argument — or that the intuition goes into what I'd argue is the wrong direction.

Pascal's Wager is sort of the most interesting of the three, because it has the clearest implications. I think i'd take it seriously if someone says they do religious thing X and Y for that reason. (Although there is a problem if X constitutes actually believing in Christianity, because you can't really choose to believe it.)

HideHide
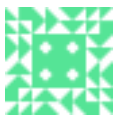
Robert Jones *says:*

June 4, 2019 at 4:08 am

I pretty much agree with all of this. In particular, Pascal's Mugging is the only one of the three mentioned problems which seriously troubles me. Even then the problem isn't that I'm convinced by it but fail to act on it: I'm *not* convinced by it. I think there's been a sleight of hand, but I haven't spotted it yet, and I find the usual critiques unconvincing. I also think it *might* be right, but then I have the difficulty that

if I assign some probability to the argument being right, I should pay the mugger. I'm also concerned that it may be illustrate something profoundly mistaken in my model of the world.

Aapje *says:*

June 4, 2019 at 7:21 am

> *I'm also concerned that it may be illustrate something profoundly mistaken in my model of the world.*

I think that the most likely mistake is the same problem with utilitarianism: an overestimation of how well one can predict the outcomes of one's actions.

If one introduces uncertainty in decision making, then vastly improbable cases with implausibly high rewards disappear in the noise.

uau *says:*

June 4, 2019 at 8:09 am

> *I pretty much agree with all of this. In particular, Pascal's Mugging is the only one of the three mentioned problems which seriously troubles me. Even then the problem isn't that I'm convinced by it but fail to act on it: I'm not convinced by it. I think there's been a sleight of hand, but I haven't spotted it yet, and I find the usual critiques unconvincing.*

Why are the critiques unconvincing? To me Pascal's wager/mugging have always seemed pretty clear-cut and obvious – they're both about failing to account for relevant alternatives.

Pascal's Wager considers the possibility of Christianity being right, but fails to consider the possibility of there being a Satan more powerful than God who eternally tortures those who worship God but leaves the rest alone, and other such alternatives. This proves the argument wrong. If someone wants to then make a modified argument and try claiming that Christianity is more likely to be right or something (as has already been done in this thread), that

can be responded to separately, but that *does not save the original argument*. If you need to resort to that, it just proves that Pascal's Wager as it's normally understood is wrong.

Pascal's Mugging similarly focuses on a single low-probability but extremely weighty possibility. But if you accept that there's a positive, even if low, probability that the mugger really might kill HUGENUM people, then it's not right to consider ONLY that one single extreme-weight possibility and implicitly assume that everything else is less weighty. If you give him money, does that encourage the mugger to commit more crimes in the future, and carry out his threat in some of those cases? Are there other people with magical powers, who could have a lot bigger impact that dwarfs anything an idiot who uses such powers for mugging could achieve? Maybe you should ignore the irrelevant mugger and concentrate on the possibility of those people whose actions would *really* matter? Maybe you should have considered this weighty possibility of magic even before you met the mugger? Basically, if you're willing to consider the possibility of such magic existing, that should shape your actions in general. That you meet a random guy who makes one particular claim about such magic, with no real evidence, is not a reason to focus only on that single possibility. Either your payoff matrix has lots of entries about possibilities with magic, or it should ignore the mugger's threat too.

Hide ↑Hide ↑
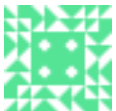
Robert Jones *says:*

June 4, 2019 at 3:26 am

The doomsday argument is not rigorous and can't be made rigorous. It's not possible to deduce a distribution for N from knowing the distribution across N items. Some prior distribution of N has been implied but not defended. I am not a randomly sampled human.

The argument must be wrong because it's independent of any empirical facts about the universe. If we lived in a universe with no observable X-risks and abundant resources, we would get the same estimate.

HideHide

b_jonas *says:*

June 4, 2019 at 4:06 am

Some people in your comment threads claim that they have signed up for cryogenic life extension. I think those people accept Pascal's Mugging and live their life according to its implication.

HideHide

An Fírinne *says:*

June 4, 2019 at 5:26 am

The simulation argument is self-defeating. If the argument is true then all of the premises (Human beings are likely to do XYZ) are rooted in observations about simulated reality as that's all we have contact with. If that is true then we have no reason to assume non-simulated reality is anything like the simulation we are living in. If this is the case then the argument is invalid.

HideHide

Murphy *says:*

June 4, 2019 at 5:38 am

Simulation of a universe seems computationally expensive to the point where unless you had some infinite source of computation it would be hard to see it as worthwhile.

But of course if you had such a resource you could provide it within the simulations practically for free.

Which would have certain implications.

https://qntm.org/responsibility

Hide ↑Hide ↑

DinoNerd *says:*

June 4, 2019 at 6:27 am

My intuition says that there are conflating factors.

If an argument is *really* about facts, rather than values, then it makes a lot more sense to be convinced by evidence. The eugenics argument (mentioned above) is about values. Ditto the never-ending abortion argument.

The more my believing (and acting on) your argument benefits you, the less credulous I should be.

And finally, the least wrong thing to think is often; "I don't know; here are the common positions" or even "I don't know, but *not* any of the following ..."

HideHide

greghb *says:*

June 4, 2019 at 7:13 am

Is there a good way to practice the skill of working through arguments? Could someone compile a book of cases? Each case really does have a totally right answer, but the true knock-down information is held out from the case materials. Instead, the case presents some arguments for and against that are meant to be as persuasive as possible. Maybe the arguments have to be drawn from sources that predate the ultimate settling of the debate. You read the arguments, pick a side (or assign probabilities), score yourself, and see if you're any good at finding truth in argument.

HideHide

NoRandomWalk *says:*

June 4, 2019 at 8:15 am

Honestly? Literally this Blog. Scott has a writing style of in many posts Making Argument A, Then Making Counterargument B, Then Revealing why A is Correct or the True Answer is C.

I often agree with his final conclusion, and when I was less smart/educated/younger would, in the process of reading a post, first believe A, then B, then C. And then feel foolish.

Now I can usually predict what Scott thinks is the correct answer, because I spot the flaws in A or B, and as a result often disagree with his final conclusion as well using the same tools.

Hide ↑Hide ↑

Spookykou *says:*

June 4, 2019 at 12:40 pm

Agreed, my desire to live under the rule of a monarch did a 180 after reading this blog

Jiro *says:*

June 4, 2019 at 7:28 am

The doomsday argument is contradicted by the self-indication assumption. This is well known by now. There's no dispute (even among experts) that mathematically, it counters the doomsday argument; the only dispute is over whether the *assumption* is valid, which isn't really the kind of expert opinion Scott is talking about.

Pascal's Mugging can be solved by noticing two things:

1) Because a larger number is more useful to frauds to fool a naive rationalist, you should discount the probability more the higher the number is. When this turned up on lesswrong people demonstrated that discounting the probability based on complexity is not enough to refute the mugging, but I am not discounting the probability based on complexity.

2) It is not possible to have a uniform distribution over an infinite set, so the distribution of possible numbers that either a fraudster or Omega can give you must have a peak. This affects the solution in obvious ways. Of course you may not know enough about this distribution to resolve the question, but the peak is probably lower for a fraudster than for a real Omega. (The fraudster is at least limited by the number of words he can say in a human lifetime, unless he points to an outside source, but the outside source is also limited....)

HideHide

jermo sapiens *says:*

June 4, 2019 at 7:30 am

This is why free speech is so important and de-platforming is so dangerous. The powerful are by definition the ones who get to decide what argument gets aired and what argument gets suppressed. Favoring the ability of platforms or governments to shut down speech they dont like is allowing the powerful (i.e., rich) to decide what is considered true. The rich and powerful already have a massive advantage in this regard, there is really no need to eliminate their modest opposition of less powerful people.

That means that racists, homophobes, conspiracy theorists, etc. get to speak. This is a small price to pay for the ability of everyone else to challenge the consensus of the powerful. Because after the racists are shut down, the censor wont pack up and leave. He'll go after 911 Truthers, people who criticize Israel (it's really easy to lump Israel criticism with racism), and even people who would like to report that their daughter is being groomed and raped by muslim gangs. And then people self-censor because if you dont you'll be in the basket of deplorables.

NoRandomWalk *says:*

It's hard, dude. Like sometimes 'the powerful' is 'the vast majority of people in society are nice and trying to coordinate exclusion of jerks from their community'.

I 100% agree with you because of I guess being pascal-wager-mugged historically by the totalitarian experiments in the 20th century, but I gotta admit that in certain cases my meta principles tell me one thing 'dont deplatform alex jones even though he totally is the reason parents of school shooting survivors suffer harassment', and then society decides to do it anyway, and on reflection it's not obvious to me that the harm to norms outweighed what I perceive as a big, local benefit.

I continue to be dogmatically opposed to any kind of censorship because I think the status quo is good enough, and the alternative can be really really bad, but I do recognize that the optimal solution (in a rawlsian veil of ignorance sense, not in the more convenient 'my views are true but sometimes powerful people may not share them' sense) may involve some small amounts of censorship and room for freedom of association by coordinated private actors leading to de facto social exclusion from anything resembling a public square.

jermo sapiens *says:*

> *Like sometimes 'the powerful' is 'the vast majority of people in society are nice and trying to coordinate exclusion of jerks from their community'.*

Not just sometimes. Most of the time that is true. I would even encourage

such behavior for private groups, but not for society at large.

Alex Jones is the perfect example. He's certifiably insane, peddles a bunch of nonsense theories, but shutting him up by force is a losing proposition in the long run. This has been discussed here. It's quite plausible that an Alex Jones type be the one who breaks an important story that no one in the respectable press would touch with a 10-foot pole.

Also, shutting him down did hurt him financially, but did not destroy him, and only increased the influence on the people who believe him. He now has millions of followers who are more convinced than ever that he is correct on everything and that the establishment fears him for speaking the truth.

With respect to the Sandy Hook thing, that was in 2012 and he was banned from social media in 2019. That's not what got him banned. And the people who believe that Sandy Hook was a hoax will believe the next Sandy Hook is a hoax with even more fervor. Banning Alex Jones will not prevent grieving parents from being harassed by idiots. Quite the contrary. You're going to have some morons on the internet being wrong about stuff, better to find a way to live with it than to waste resources trying to police opinions online.

The thing is with violation of norms is that the bad consequences may come much later. It's naive to think that if within 2 years of an event, all consequences of that event have played out. Or replace 2 years with 10, 20, 100…

We know the cost of letting the powerful control what the truth is. They can be wrong or they can even be lying. Whatever they do, we can be certain that what they peddle as the truth is going to be the proposition that helps them maintain their grip on power. Accepting a ban on Alex Jones today, however odious Jones may be, is accepting a ban on a slightly less Alex-Jonesy guy tomorrow, and on and on, until you ban someone for saying something like "men arent women".

Letting the elite have a monopoly on information is what censorship does. And whatever benefit derived from the banning of Alex Jones does not even come close to balancing that.

sharper13 *says:*

I think sometimes we have a tendency to quash those we don't like/don't agree with by blaming them for the "consequences" of their views and trying to associate them with their craziest supporters, rather than just solving the problems we blame them for.

For example, I don't support (nor really know all that much about, as I've never bothered to really listen to him, not appealing to me) Alex Jones, but the obvious solution to the problem of Alex Jones "is the reason parents of school shooting survivors suffer harassment" is to very publicly take action against the people actually doing the thing which is bad, i.e. harassing those parents.

Now perhaps you have some counter-argument that Alex Jones is able to control the actions of others in order to make them do bad things, but I don't think that's the default assumption. The default assumption IMHO should be that the people actually doing the bad things should be held responsible for their actions, rather than blaming someone else for speech which "convinced" them of a premise which led to them thinking the bad thing was a good idea. Note that this is different than someone literally saying, "You should go do this bad thing because it's really a good thing."

If you can credibly make the argument "Listening to Alex Jones will land you in jail because his arguments will lead you to do bad things", then maybe that will reduce his number of listeners. If the argument is rather "This guy is unstable and he last listened to Alex Jones before he harassed someone, so let's go after Alex Jones.", then the flaw there seems obvious, which is that the same guy will probably just listen to someone else and go harass someone anyway.

TL;DR Both are in the group of people I disagree with, but I don't blame Bernie Sanders for his supporter shooting up the GOP congressional baseball team, so why would I blame Alex Jones for his supporter harassing parents of dead children? People are responsible for their own actions.

(Note: I'm assuming for the sake of this argument that Alex Jones never literally said people should harass those parents, I don't actually listen to him, so I wouldn't know that for sure.)

Hide ↑Hide ↑

NoRandomWalk *says:*

June 5, 2019 at 11:26 am

@sharper13
I don't listen to his shows, but I've listened to him when he goes on other shows.

As far as I can tell he's a conspiracy theorist who believes the conspiracies he advocates (more or less, unless he's fighting for custody of his kids in which he says it's all an act, which is understandable).
In a couple rare instances, he got the conspiracy theories completely right according to people I respect but I don't know the details of.

I totally agree that, morally, neither Bernie Sanders nor Alex Jones are ethically responsible for what their supporters do.

I think also that, mechanically, it is impractical (and undesirable) to make it illegal to harass people.

It's not obvious to me that the harassers would have harassed someone else if not for Alex Jones. Some people are like that, some people would have just stayed normal (would Germany have tried … stuff …. without …. some guy).

I'm just saying it's quite possible deplatforming alex jones actually helps people experience less harassment.
And that it's a much more effective way to reduce harassment with few side effects than trying to find a legislative solution that 'makes harassment illegal but doesn't infringe on free speech rights'

Aapje *says:*

I once decided to follow up on [an article](#) that I found on Infowars (after googling), which intrigued me, and it did turn out to be true on the whole. It was the claim that the BLM founders admire a radical black activist, Assata Shakur, who was convicted of first degree murder of a cop; and that BLM recites an Assata Shakur quote at events.

My own research found three separate interviews on far left websites, with the three different BLM founders, where they call Assata Shakur their inspiration. In one of the interviews, a founder says that BLM recites the quote. Googling pictures of BLM protests also turned up references to Assata Shakur and her writing, on placards and clothing carried/worn by protesters.

I wouldn't call this claim by Infowars a conspiracy theory, because they didn't allege any secrecy or extreme & unlikely coordination between people. The idea that an activist organisation would pick a inspirational hero and use their writing in their activism is far from peculiar.

If anything, it increased my belief in a 'conspiracy theory' that is not mentioned in the article: the idea that (center) left-wing media tends to ignore it when people/groups on the left treats radicals with a history of violence as heroes.

I also updated a little towards the idea that Infowars is not completely insane.

Matt M *says:*

if I know that a false argument sounds just as convincing as a true argument

> *argument convincingness provides no evidence either way. I should ignore it and*
> *stick with my prior.*

I went through a primitive version of this when I was in high school and started getting involved into politics. Came to the conclusion that politically biased journalism (which roughly maps to all journalism) can *always* find certain statistics and certain spin to "prove" whatever they want.

There are an approximately infinite amount of reasonable sounding sources that "prove" socialized medicine results in superior health outcomes and general wellbeing, as well as an approximately infinite amount of reasonable sounding sources that "prove" it results in a totalitarian hellscape.

So I decided to not worry about determining the "facts" of any of these issues much at all, and to just go with what felt right to me based on my own preferences. Whether that worked out for the best is… well, opinions may vary on that, too…

HideHide
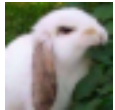
Dan L *says:*

June 4, 2019 at 11:01 am

It's easy to lie with statistics. It's easier to lie without them. The delta is useful.

More rhetorically, it's interesting to see this be argued as a defense in the face of a "proof" the holder dislikes. One would think that if it were true, its adherents would have little trouble countering appropriately, secure in the knowledge that their interlocutor has denied themselves a similar retreat. If they fail to do so, maybe that's evidence against.

Hide ↑Hide ↑

Deiseach *says:*

June 4, 2019 at 8:05 am

But to these I'd add that a sufficiently smart engineer has never been burned by arguments above his skill level before, has never had any reason to develop epistemic learned helplessness. If Osama comes up to him with a really good argument for terrorism, he thinks "Oh, there's a good argument for terrorism. I guess I should become a terrorist," as opposed to "Arguments? You can prove anything with arguments. I'll just stay right here and not blow

myself up.

It's also that one man's terrorist is another man's freedom fighter or Resistance operative. What makes someone who blew up bridges and ambushed soldiers a hero in one instance and a terrorist in another? "Oh, but the soldiers they ambushed were the German army during the Second World War in Occupied France, and the other lot hijacked planes and flew them into American buildings!" Yes, and from the German point of view, the Resistance were terrorists and from the hijackers point of view they were soldiers for truth fighting a dreadful enemy. Had the American Revolution been crushed, today the [Boston militia](#) might be remembered as traitors, rebels and terrorists instead of heroic soldiers in the army of liberty.

[Michael Collins](#) – terrorist, freedom fighter, statesman? How about all three at different times? Maybe the Irish equivalent of bin Laden doesn't walk up to a Catholic young man in Derry and go "Hey, wanna be a terrorist? Here's three great reasons why terrorism is cool!", they say in the aftermath of [Bloody Sunday](#) "Want to protect your family and loved ones and fight back against a murderous occupying force?"

And then [one day](#) you put down the Armalite, rely on the ballot box, and make [compromises](#) with your deadly enemies.

This is not to argue against Scott's point, it's to show that it's not an easy clear-cut choice between "obviously bad and wicked" and "obviously true and good" arguments.

[HideHide](#)

kokotajlod@gmail.com *says:*

June 4, 2019 at 8:43 am

I notice lots of people making what they think are coherent arguments against the simulation argument, the doomsday argument, and Pascal's Mugging.

Here is my [sequence on Pascal's Mugging](#) and here is my [MA thesis on anthropics](#).

I hereby offer a $100 bounty, to be divided evenly between the first 3 people to convince me that I'm wrong about some particular thing I said in one of those documents. I look forward to talking with you, whoever you are!

Note: The MA thesis doesn't spend much time talking about the simulation argument or the doomsday argument; it talks about the different rules of reasoning that lead to the arguments

going through or not. But if you accept the main claim of my thesis–that SSA is best–then the Doomsday argument follows. The Simulation argument, meanwhile, is more robust: The only anthropic assumption it uses is the bland indifference principle, which is pretty universally endorsed. The simulation argument uses other assumptions–such as the assumption that simulated humans are conscious–but it is pretty good about clearly acknowledging its assumptions. I'm happy to extend the bounty to also cover people convincing me that the simulation argument is wrong.

Another note: I currently think that updatelessness/ADT, which I did not consider in my thesis, may be correct. So merely pointing this out to me won't count. But if you can convince me that it is correct I'll count that. In general I'm happy to pay for major updates to my views on topics I consider important. 🙂

HideHide

Deiseach *says:*

June 4, 2019 at 9:01 am

I haven't anything anywhere near approaching the maths to evaluate probabilities in Pascal's Mugger, but dragging it down to boring concrete examples, a real Mugger won't threaten to kill X infinity power number of people if you don't give them five bucks (or that X infinity power people will die), they threaten to kill/harm *you*. Imminent personal bodily harm is a heck of a lot more persuasive.

If somebody tried the "I'm from an alternate dimension and I desperately need a fiver to save X infinity number of lives" approach, I might give it to them (if I had it to spare) not because "Well, sure the tiny probability works out!" but because the sheer cheekiness of the approach amused me.

Hide ↑Hide ↑

Conrad Honcho *says:*

June 4, 2019 at 9:31 am

I'm more interested in seeing the sorts of contrarians who will take you up on that argue over what to do with the leftover penny when dividing $100 evenly three ways.

Hide ↑Hide ↑

Spookykou *says:*

June 4, 2019 at 12:37 pm

My first thought

Deiseach *says:*

what to do with the leftover penny when dividing $100 evenly three ways

Flip it to see who gets it. If two call the same (heads or tails) and it comes up that result (heads or tails) then flip it between them to see who gets it. If everybody is a smartarse and calls out the same (heads or tails), then the neutral referee gets to keep the penny.

uau *says:*

I briefly skimmed your writing about Pascal's Mugging. You seem to agree with my view that Pascal's Mugging itself is basically resolved by considering the alternatives omitted from the problem statement – if you are generally willing to accept the possibility that there are ways to save HUGENUM people, then there will be better ways to do that than giving money to a likely-lying mugger; the error is in considering the small probability of saving people that way, while failing to consider saving even more people in other ways.

As for the more general case of extreme positive or negative payouts, I think the most reasonable approach is to say that if any "extreme" events are plausible, creating a powerful and well-functioning civilization has the best chance of achieving the good ones in the future (if an interdimensional traveler on a quest to save HUGENUM people appears, such a civilization will give the best chance of being able to help). I disagree somewhat about your related point about "cluelessness": while it may be true that all actions do not have exactly the same probability of pleasing an all-powerful genie, I think it's reasonable to say that rather than focus on genie-pleasing *now*, it's better to advance knowledge and technology. P(concentrating all resources at current best effort at pleasing genies convinces one to convert the world into a paradise) < P(after concentrating on general development that finally allows scientific study of genies, people convince one to convert the world into a paradise).

Edit to add elaboration:

In general, I think the "cluelessness" argument works better than you give it credit for. We do not have reason to believe that any particular action has really extreme utility. That there could be a minuscule probability of something having huge utility can be countered by saying that getting knowledge and improving general abilities is more important. "It's *possible* we could get huge utility by doing X, essentially by chance" generally implies "there are ways to get absurdly huge utility, improving our general capabilities increases our chances of achieving that".

Hide ↑Hide ↑

kokotajlod@gmail.com *says:*

June 4, 2019 at 12:11 pm

Thanks for the reply!

I agree with your first paragraph. As for your second paragraph, notice that the decision theory you are implicitly using is subtly different from expected utility maximization. Expected utility maximization breaks down in realistic cases, because there are infinitely many possibilities and they don't sum up to anything (every action has undefined expected utility!). So instead you are saying we should do the thing that has the highest probability of yielding a hugely good outcome. How good is hugely good? Infinity? Which infinity then?

At any rate, I am not as optimistic as you that this new method (maximize probability of extremely good outcome) adds up to normality. For example, suppose we do some physics and decide that because of entropy and the speed of light there is a finite limit to how much good we can do. On your view, we should then spend all of our resources checking and re-doing our calculations so that we can increase the chance of getting infinite utility anyway. But this is absurd–in this situation we should give up on infinity and settle for a finite utopia.

Hide ↑Hide ↑

uau *says:*

June 4, 2019 at 12:35 pm

> *For example, suppose we do some physics and decide that because of entropy and the speed of light there is a finite limit to how much good we can do. On your view, we should then spend all of our resources checking and re-doing our calculations so that we can increase the chance of getting infinite utility anyway.*

I'm not convinced that this would follow. Especially not in any silly "do the same calculation over and over again" sense. Once you're reasonably certain that the calculation has no obvious mistakes, I think a better strategy would be to assume that you need new ideas to change some of the underlying assumptions of your calculation, and go back to advancing science in general.

Hide ↑Hide ↑

kokotajlod@gmail.com *says:*

June 4, 2019 at 1:30 pm

OK, sure–but the general point still stands; at some point we need to stop advancing science in general and start building our happy utopia full of people only some of whom are optimized scientists.

Hide ↑Hide ↑

uau *says:*

June 4, 2019 at 2:08 pm

> *at some point we need to stop advancing science in general and start building our happy utopia*

Do we? Sure, it sounds like a reasonable thing to do. But I don't see opposite option as fundamentally inconsistent. Practicing strict utilitarianism where you really value the lives of strangers arbitrarily high as long as there are many enough strangers already seems kind of utopian. Would refusing to settle for limited good be somehow more inherently wrong?

kokotajlod@gmail.com *says:*

I think we do, yes. I think that fanatically pursuing long-shot attempts to get infinite good is a bad idea; I think there comes a point where we should devote at least some resources to something else instead. Like feeding the hungry.

I'm not a classical utilitarian; if I was, I would have an unbounded utility function and I probably would agree that the best thing to do is maximize probability of infinite reward. (But which infinity?)

DeadAtheist *says:*

> *Bostrom's simulation argument, the anthropic doomsday argument, Pascal's Mugging – I've never heard anyone give a coherent argument against any of these*

At first I didn't believe it, but wow, so many bad arguments are floating around. I will try to add some good ones. Not to waste everyone's time, one line debunks:

Bostrom's argument: Statement 2 is correct. Can you really imagine a programmer spending hours upon hours, trying to get cancer victims to feel the correct amount of suffering?

Doomsday argument is *a* way to get prior probabilities of doom. Nothing else. It's not inescapable, this probability should update like any other.

Pascal's Mugging is just too computationally costly for human brains to be any better than chance, so might as well keep 5$.

kokotajlod@gmail.com *says:*

Thanks!

Your objection to Bostrom's argument isn't really an objection, since his argument is for a triple-disjunct. But anyhow, no I find it hard to imagine a human programmer doing that, but I find it easy to imagine an AI programmer doing that.

Similarly, you aren't really disagreeing with the Doomsday argument–you are just pointing out (correctly) that even if you accept the argument you still have to update the probabilities based on your evidence afterwards. The people who write about the Doomsday argument agree with this.

I'm not sure I understand your objection to Pascal's Mugging. If you are saying that we are clueless as to whether giving the money or withholding the money is more likely to lead to hugely good outcomes… I'd say we aren't clueless, we can be pretty confident that withholding the money is more likely to lead to hugely good outcomes! I think there are good objections to Pascal's Mugging taken as an argument that you should pay the mugger, but taken as a paradox or puzzle for expected utility maximizers I think it's pretty solid. (Even if you shouldn't pay the mugger, it's a problem that the expected utility of every action is undefined, and that your behavior gets more and more crazy and erratic as you try to approximate the impossible ideal by considering more and more possibilities more and more rigorously)

Hide ↑Hide ↑

DeadAtheist *says:*

June 5, 2019 at 12:07 am

Ok, to clarify: I was specifically responding to Scott's concern that he never saw debunks of these arguments, yet smart people he knows do not live their day-to-day lives according to them. So I was not debunking them completely – only in the forms that make you change your life! Otherwise they have their merits.

> *Your objection to Bostrom's argument isn't really an objection, since his argument is for a triple-disjunct. But anyhow, no I find it hard to imagine a human programmer doing that, but I find it easy to imagine an AI programmer doing that.*

Yep, Bostrom's argument is technically correct, in the same way I would be correct saying "Either I'm dumb, or the Earth is flat". Argument from morality clearly shows that specifically ancestral simulations are unlikely, and that means you have no reason to, e. g. seek ways to precommit super-hard to saving everyone stuck in the simulation then you wake up (if it turns out you visited simulation for recreational use without your memories) or to move stars to write "ALZHEIMER'S SUCKS, NO REALLY, YES EVEN COMPARED TO EXISTENTIAL HORRORS OF POST-SCARCITY SOCIETY" in the night sky. Argument from morality *doesn't* show, as you noticed, that we can't be in non-ancestral simulation, but that hypothesis is not supported by anything solid and doesn't imply we should be doing something about it even if true.

> *even if you accept the argument you still have to update the probabilities based on your evidence afterwards. The people who write about the Doomsday argument agree with this.*

Really? When what's the big deal about it?! (Apart from catchy name). Why do people try to refute it so hard, if all it does is provide one of many ways to set your subjective priors? We have tons of evidence about possible end of humanity, any reasonable prior should be irrelevant.

> *I'm not sure I understand your objection to Pascal's Mugging.*

Pascal's Mugging was invented as an example to show that without leverage penalty (or something similar and hopefully better) expected utility diverge, making coherent decision-making impossible. Correct so far. What it means for *you* is that if you are the kind of person who can contribute to the decision theory, you should, because right now it's broken. It doesn't mean you should give money to muggers any more than Zeno's paradox means you can literally never be physically shot, so you should go have a walk on active shooting range.

Hide ↑Hide ↑

kokotajlod@gmail.com *says:*

re: Pascal's Mugging: Agreed.

re: Doomsday: Well, many people feel intuitively that the DA is wrong even if it is just one factor among many that contributes to your credence. So that's why they argue about it. Moreover, it's a pretty weighty factor; I'd guess that for most people it makes the difference between thinking we'll probably die and thinking we'll probably survive. I'd love to be convinced otherwise; maybe some calculations are in order?

re: Simulations. I don't think the definition of "ancestor simulation" requires human programmers. If it does, then sure, fine, but just change the definition to be a bit more broad and it goes through.

Besides, Bostrom's triple-disjunct is waaaaaaay more interesting and powerful than "either I'm dumb or the earth is flat." Notably missing from the disjunct is the view that most people hold, which is that there's a non-negligible chance that we are not in a simulation yet will also not die or create ancestor simulations.

(Also, by the way, I think the argument from morality is pretty weak too–it is way too optimistic about moral progress and simultaneously too unimaginative about ways that creating simulations might actually turn out to be justified.)

Hide ↑Hide ↑

DeadAtheist *says:*

June 5, 2019 at 12:45 pm

re: Doomsday: Ok, it seems that i was wrong in understanding of other people's position on it. But I would maintain that it shouldn't be "a pretty weighty factor". Physics, biology, climate science, AI research and so on give us detailed model of our world and we can predict humanity's fate with that model. For an example:

A = humanity is killed by rogue AI in no more than 100 years.
B = AI alignment field is currently on fire

P(A) = 0.01 by Doomsday Argument

P(B|A) = 0.95

P(B|~A) = 0.1

P(A|B)= P(B|A) * P(A) / (P(B|A) * P(A) + P(B|~A) *

P(~A)) = 8.75%

Just one factor changes odds a lot. And there are thousands or relevant factors, so even if you don't agree with my numbers (I don't stand by them), they must compound.

re: Simulations: It's even harder to imagine friendly AI doing it. If it's an AI not aligned with human values, in what sense is it ancestral simulation?

> *Notably missing from the disjunct is the view that most people hold, which is that there's a non-negligible chance that we are not in a simulation yet will also not die or create ancestor simulations.*

Then I'm not a typical human in that I don't find an idea of ancestral simulations all that intuitive. You are a human in year 3569, your options are almost unlimited, you have vast amounts of computing power, and what do you do with it? Delude yourself into thinking you are minimum-wage worker in 2019-Detroit with no prospects in life and any day a carbon copy of another, because that's the most exiting option apparently. Really? And any scientific application of such a simulation can be done without real minds inside.

> *(Also, by the way, I think the argument from morality is pretty weak too–it is way too optimistic about moral progress and simultaneously too unimaginative about ways that creating simulations might actually turn*

out to be justified.)

Um… we already realize that causing Holocaust is bad and are unlikely to change our mind about it. There might be reasons for me to be wrong, but I can't take into account info that doesn't yet exist.

Hide ⬆Hide ⬆

kokotajlod@gmail.com *says:*

June 5, 2019 at 1:32 pm

re: Doomsday: You give a nice example in which
$P(A) = 0.01$
$P(B|A) = 0.95$
$P(B|{\sim}A) = 0.1$
The size of this update is about one order of magnitude.

If the Doomsday argument is valid, then we get something like:
$P(OurBirthRank=10^{11}|Doom) = 10^{-11}$
$P(OurBirthRank=10^{11}|{\sim}Doom) = 10^{-60}$
So when we find out that our birth rank is 10^{11} or so, we make a MASSIVE update towards Doom–about fifty times the size of the example you gave.

If there are indeed thousands of relevant factors, then yeah you might still end up disbelieving in doom. But the Doomsday Argument is arguably the biggest single factor to think about.

(Disclaimer: I'm playing fast and loose with numbers here and I have a sneaking suspicion I may be making some mistake)

re: Simulations: I thought the definition of an ancestral simulation was one made to simulate your ancestors. An unfriendly AI still has us as ancestors–not biologically, but still, its chain of creators stretches back to us

I'm not sure I follow the point you made about not being like most people. It sounds like you agree with most people–you think that we aren't in a simulation, and also we aren't doomed, and also there's a non-negligible chance that simulations will be created in the future. Or are you saying the chance of creating simulations is negligible, for moral reasons? Again, I think you are being way too optimistic.

I have a few ideas in mind for why it might be a good idea to create simulations (disclaimer: I am not convinced by them, not at all, but it's enough to make me think the topic merits further reflection) but I don't really want to talk about them now. If you are interested, ask me some other day.

Hide ↑Hide ↑

nathan_young *says:*

June 4, 2019 at 8:51 am

Thank you for writing this. It contains a lot of intellectual humility. This seems to to assume that good arguments are indistinguishable from bad ones. I suggest there is a better way of navigating the issue.

1. Try and understand others arguments to find the flaws in your own. It's far more useful to see where your own argument falls than try to work out if another succeeds, but often others are better at finding that than I am. I think that the "Feynman Method" – you don't understand an idea until you can explain it – is also useful here. Often other people's wonderful ideas sound much less good when you try to explain them to someone else.

2. Arguments without rigid, testable mechanisms are just rules of thumb. If someone suggests something is a certain way, but doesn't say why and how it could be found not to be the case, it may be useful but I wouldn't take it to the heart of my epistemology. When you have a rigid, testable mechanism, you can find any number of tests to see if it holds up.

These two ideas can be combined in the notion that rather than trying to believe truths we

should try and avoid believing falsehoods. In relation to your post, this means I try to understand people's theories (especially when they provide mechanisms for why they are the case) and listen when they try to critique my own.

Thanks again, interested to hear others thoughts.

HideHide

eric23 *says:*

June 4, 2019 at 9:40 am

I have to say this reads as a repudiation of the entire rationalism movement. If the correct response to a convincing argument is so often "Arguments? You can prove anything with arguments" then does all that mental training have a net positive or negative effect? And even if it has a net positive effect for the hyper-intellectual fringe who frequent SSC and LessWrong, does it have a net positive effect for the average plumber or Uber driver?

HideHide

Bugmaster *says:*

June 4, 2019 at 10:00 am

According to this article, the training has no effect whatsoever, because even a trained person such as Scott would simply go with his gut all of the time.

Hide ↑Hide ↑

eric23 *says:*

June 4, 2019 at 10:19 am

That too is a repudiation!

Hide ↑Hide ↑

rahien.din *says:*

June 4, 2019 at 10:11 am

People aren't convinced by logic and evidence. People are drawn to/by a feeling of shared belief or common ground. (This is why steelmanning is an effective debate tactic, and prophets get murdered.)

As they interact with other people, every person offers their partner some measure of common ground. Epistemic learned helplessness, at heart, is a constriction or restriction of that offer.

Logic appears later in the process – such as when you justify yourself by saying "This is the correct Bayesian action." You are right, of course – but why are you saying it and to whom do you make that declaration?

HideHide

The Nybbler *says:*

June 4, 2019 at 10:11 am

The whole engineers-as-terrorists thing doesn't really hold up. Especially if you say it applies more to "sufficiently smart" engineers. Only one of the 9/11 hijackers was an engineer. One other had some technical training and was a lousy student. Three were law students, who certainly ought to know that you can put forth a convincing argument for just about anything. The 2007 attack on the Glasgow airport involved an engineer... but also a doctor.

HideHide

Watchman *says:*

June 4, 2019 at 10:53 am

Is this post not giving too much credence to the power an argument has? How often are peoples' minds changed by a single argument? My core beliefs have changed over time, but I can't think of one belief that has ever changed because of one argument or fact. Instead they've slowly morphed in reaction to a variety of arguments, facts and experiences. The only belief I can firmly identify a source for comes from an authority figure's decree: when young I grew up in effectively an all-white environment, so was starting to use language labelling people with other skin colours as other; my dad calmly explained they were people just like us and colour was irrelevant, and this has stuck with me. Most people I know well are the same: they have firm views learnt in childhood and other views slowly shifting around but not being suddenly changed by any one thing.

To take an example from religion, an area where one would expect to find the idea of the killer argument causing change, the treatment of conversion to Christianity in early-Christian literature is significant. The focus is on teaching and good works, and in Roman-period works often on being a living example. The sermon that converts a ruler or a crowd is a miracle in these works: an exceptional event showing God's favour. It is the same as a healing or the defeat of a magician as a way of showing the power of the true religion (in the eyes of the ancient writer), and not something that was expected to happen. Saints and missionaries instead talked to people and lived amongst them and conversion to Christianity was a process

instead talked to people and lived amongst them and conversion to Christianity was a process not a moment.

That there is no sudden flash of light, no moment of conversion, explains why people encountering strong arguments tend not to adapt to them. Another way of looking at this is how we educate people: we don't expose them to arguments and wait for the eureka moment, but rather direct and sometimes dictate; we let learners experience and encourage peer interaction. No education system has ever based itself just on the power of arguments to change minds, which suggests that there is a long-standing pedagogical awareness that arguments do not work to change minds, whereas a combination of decree, experience and argument does so.

This is not to say that epistemic learned helplessness is not a thing, but rather to say that the thing that it is is normal in humans, an innate reluctance to change in the face of a new argument, however convincing. The argument here assumes arguments should change minds, but there is actually very little evidence that this is the case.

HideHide

llamagirl *says:*

June 4, 2019 at 12:27 pm

My rules for epistemic humility:

1) Separate *decision-making under uncertainty* from *epistemology*. We have to make decisions on imperfect information all the time, but we can stop our minds from implicitly assuming that decision entails sanctifying the premises that drove us.

2) Avoid belief-formation *as the default*. Do I really need to have a stance on subject X? Or can I just remain undecided and mildly curious?

3) Avoid certainty and instead think in terms of likelihoods. I find it helpful to visualize degrees of belief as shades of gray.

4) Consider external validity in addition to internal validity. What impact might *unknown unknowns* have on the truthiness of my belief, in addition to the *known unknowns*?

5) Consider the reasons why an argument may be convincing that have nothing to do with its truth value. Is it unfalsifiable in structure? Does disbelief come at a social cost? Does belief offer

rewards? Would disbelief necessitate significant upheaval of my world-model?

6) Never dismiss proponents or opponents of an argument as [murderists](). Ascribing bad intent takes very little effort. Refusing to dismiss opponents as evil drives a certain open-mindedness and curiosity – e.g. how is it that good and loving humans voted for Trump? There are interesting answers, but first one has to eschew the cognitively easy route of evil-paper-dolling.

7) Take a hard look at beliefs that appear to be exempt from logical or empirical scrutiny, e.g. "self-evident" claims or increasingly meta beliefs (there's no such thing as an epistemological free lunch, not even for a belief about beliefs!)

HideHide

SEE *says:*

June 4, 2019 at 12:30 pm

One time, just one time, did I ever convince my sister to do a deal in Monopoly.

From then on, she assumed that whatever my argument was, no matter how favorable the deal looked, it was going to work to my advantage. Which she explicitly told me when I asked why she wouldn't take a deal I was offering her.

HideHide

Randy M *says:*

June 4, 2019 at 12:41 pm

Your sister is right, at least in two player. Any deal you offer, or even accept, should give you an advantage over her. If you are better at understanding the game state, she should never trade with you.
That's different if there are three or more players. Then the advantage goes to whoever trades the most, even to the point of slight net disadvantage, so long as each deal nets them an improved situation compared to the uninvolved parties.

Hide ↑Hide ↑

Spookykou *says:*

June 4, 2019 at 1:01 pm

[epistemic status: I had a family member who came to a similar conclusion and it did result in an unwinnable game and in the face of this they stilled

refused!]

I think the sister is almost always wrong, baring a seriously poor understanding of the game, or a desire simply not to lose rather than to win. If no trading happens it is possible that nobody gets a monopoly and then the money from passing Go is greater than the cost of undeveloped rents and the game never ends. More generally, you get a monopoly and I get a monopoly, whoever gets the better one gives some money, or similarly structured trades should not be hard to come up with, and the marginal advantage you might get, by not actually paying me the difference in the expected value from the rents of our monopolies as a bonus is almost always going to be outdone by luck of the dice.

Hide ↑Hide ↑

Matt M *says:*

June 4, 2019 at 1:12 pm

I mean, Monopoly specifically is a poorly designed game in the sense that it often leads to ridiculously long stalemates, such that the "winner" is frequently the person who most enjoys playing monopoly and/or doesn't have anything better to do; because the other player ultimately concedes out of a desire to stop playing a pointless game and go do something else.

Hide ↑Hide ↑

Edward Scizorhands *says:*

June 4, 2019 at 1:34 pm

With more than 2 players, you don't get stalemates, unless people refuse to trade, which they shouldn't be refusing to do. It is literally called "the fast-dealing property trading game." The people who trade the most will win.

There are problems with Monopoly that it wouldn't have if it were invented in the past 10 years.[1] But a lot of the problems are because people don't play by the rules, or refuse to play the game. Risk would be bad if everyone refused to attack and wondered why it took so long.

[1] There should be some kind of victory condition that doesn't require killing other players one-by-one, which is also a problem that classic Risk has. There should be some mechanism that forces players to not stalemate themselves, like a chance of getting destroyed if you don't trade. It is hard to fix it as a 1-on-1 game, but Risk isn't very fun 1-on-1 either.

Hide ↑Hide ↑

The Nybbler *says:*

June 4, 2019 at 2:58 pm

> *There should be some kind of victory condition that doesn't require killing other players one-by-one*

Unless you mean you can eliminate a bunch of them at once, it wouldn't be Monopoly without that.

Hide ↑Hide ↑

moonfirestorm *says:*

June 5, 2019 at 6:07 am

> *There are problems with Monopoly that it wouldn't have if it were invented in the past 10 years.*

Remember that the game was originally created to make a political point rather than to be an enjoyable game. It was deliberately supposed to be unpleasant, to say "see? This is what happens when we let people concentrate land in private monopolies! Pretty awful, huh?"

Wikipedia says there was another set of rules designed to reward cooperation. I'm curious what those were and whether they could be applied to Monopoly in its current form. Some quick googling wasn't able to turn them up.

Randy M *says:*

If you consider staying in the game a failure state–an understandable opinion–you are right. If you want to avoid losing, I still think she is right, barring the large caveat about multiplayer– and of course it should only be played multiplayer.

Matt M *says:*

In a strictly technical sense, the only way one "wins" Monopoly is by being the last person to "avoid losing" (going bankrupt). The two terms are functionally equivalent.

The game would be different (possibly a lot better) if instead, the winner was "the first person to make $X in profit" or something like that.

Matt M *says:*

Agreed.

I play fantasy sports, and reject most proposed trades based on similar logic.

"This player wouldn't be offering me this trade unless they thought it benefitted them. Given that we're participating in a zero-sum contest, anything that benefits them, harms me. Ergo, I should only accept this trade if I am quite confident that I am smarter than they are and that it will benefit me more than them. But the fact that they proposed this specific trade among all possible trades highly implies this is unlikely to be true."

sharper13 *says:*

June 4, 2019 at 10:40 pm

@Matt M:

Does your philosophy on fantasy sports trades admit to the possibility that you could both benefit from a trade, making yourselves both better at the expense of your mutual opponents? That if trades tend to be mutually beneficial, those who trade the most will tend to win the most?

I confess, I don't know many fantasy sports with trading involved where it is only a two person game...

Have you ever played Cataan?

Hide ↑Hide ↑

Matt M *says:*

June 5, 2019 at 6:43 am

Actually yes. I accepted a trade proposed by someone else this season, in fact, but with great hesitation, and with virtually all other circumstances aligning in favor of the trade, such as:

At the time, I was quite far behind in the standings and drastic/risky action was warranted in an attempt to salvage my season.

I was trading away a player at a position I was strong in, to gain one in a position I was weak in.

And while it's true that the short-run effect of the trade is "We both get better at the expense of everyone else," in the long run, it remains true that the league structure rewards a single and solitary champion. That in order to truly "win", I will, eventually, have to defeat my trading partner.

(I briefly played Cataan several years ago, but not enough to really remember the rules right now)

Hide ↑Hide ↑

SEE *says:*

June 5, 2019 at 1:56 pm

Of course she'd be right in two-player. Which is why playing two-player is stupid and boring (as in, more so than Monopoly with three or more); if there weren't three, we didn't play. We have a brother.

As it happens, the trade that she felt "burned" her was me giving her Water Works and Boardwalk in exchange for Tennessee Avenue — giving her the utility duopoly and the dark blue monopoly (Boardwalk, Park Place), me the orange monopoly (Tennessee Avenue, St. James Place, New York Avenue).

She was excited at getting the highest-value monopoly in the game and extra property to boot. She was sure the deal was tilted in her favor, and explained my willingness by the fact that it still got me advantage over our brother.

But the most common starting position in the game is Jail. The oranges are hit on 6, 8, and 9, or one third of doubles-out-of-jail rolls and almost 40% of paid-then-roll outs. On the other hand, the "Go to Jail" space filters out people on the way around the board to Park Place/Boardwalk.

My brother accordingly eventually bankrupted on a roll out of Jail, and with all the cash and property I gained from that, she was doomed.

Hide ↑Hide ↑

wanda_tinasky *says:*

June 4, 2019 at 12:49 pm

You've never heard a coherent rebuttal of the Doomsday Argument? Are you serious? How about: we're not a random sampling from a preexisting cohort of all humans. Or that information doesn't travel backwards in time. If we were all born with random numbers on our foreheads that seemed to follow a probability distribution, maybe DA would have a point … but we're not. Self-indication gives absolutely no information about future events.

Imagine there's an enormous gumball machine, and we're supposed to estimate the number of gumballs by observing what comes out. If the gumballs were given serial numbers before they were added to the machine (as German Tanks might be), then our observations about which balls came out would give us an estimate about the total population. But if the gumballs have no numbers on them, then there's no information to analyze. Guess what: humans aren't born with serial numbers!

Seriously, if any of that is at all confusing, just say so. I'll happily point out the flaw in any of the retarded metaphors that DA-ers like to use. DA is my #1 intellectual pet peeve.

I've never heard a coherent argument for the DA that didn't immediately betray the author's mathematical naivety. Its continued existence within the philosophical literature is one of the main reasons I consider modern academic philosophy to be largely fraudulent, and why I never take anything Nick Bostrom says seriously.

HideHide

I Write Bugs Not Outages *says:*

June 5, 2019 at 12:53 am

The assumption is that, using data from archaeology, genomics, etc., one can reasonably accurately estimate the number of humans who lived before a certain time. Thus, a human's "serial number" can be approximated well enough for the argument to work.

Hide ↑Hide ↑

wanda_tinasky *says:*

June 5, 2019 at 1:11 pm

But that serial number bears no relation to a larger parent population. It depends only on the number of people already born, with no influence whatsoever from the number of people who will be born.

Take the gumball analogy. You can write numbers on the balls as they come out. What does that tell you about the number of ball still in the machine? Nothing! How can you possibly think otherwise?

Hide ↑Hide ↑

uau *says:*

June 5, 2019 at 5:12 pm

But if you can select a random ball, its serial *does* convey information. I'll write an larger analysis in a top level post.

Hide ↑Hide ↑

wanda_tinasky *says:*

June 6, 2019 at 6:35 pm

Of course, a random sampling followed by observation conveys information. But being born doesn't sample randomly from the future timeline. It samples ordinally. That's a crucial difference.

Hide ↑Hide ↑

I Write Bugs Not Outages *says:*

June 5, 2019 at 7:06 pm

So in German tank terms, your objection is that people are trying to use serial numbers to estimate the all-time number of tanks that will be produced while the war is still going on.

I like the gumball analogy: someone observes that N gumballs have been removed, and then declares that the most recently removed one has a 90% chance of being in the last 90% of gumballs. It helps to reveal the hidden assumption of a sort of scale-invariance property of the prior probability distribution for the number of gumballs: that for some/any k, the kth percentile of the part of the distribution with $n > N$ is directly proportional to N. (A probably density function like $f(n) = \exp(-an)$ satisfies this; are there any others?)

Hide ↑Hide ↑

morris39 *says:*

June 4, 2019 at 1:19 pm

Engineers are taught differential equations. These equations correlate variables but only within

definite boundary conditions and provide a useful answer if given the starting condition, otherwise the relationship is not useful or fails to give an answer. Those that pay attention later realize that in practical human situations context is as important as the correlation. In practical work, engineers are optimizers choosing a particular compromise solution based on the particular goal. To say you have heard good arguments for a tendency to black or white suggests that maybe you did not evaluate those arguments well enough. I would say that black/white thinking is more likely in arguments based on ideology, i.e, no compromise. Your point about the usefulness of logic is valid but only when you exceed the limits of your knowledge. A separate but correlated issue is the limit of human knowledge, very much overstated (I think willfully). Maybe you will consider an essay in the future?

HideHide

michael_b *says:*

June 4, 2019 at 3:10 pm

Agree. I think my best intellectual breakthrough has been to refuse to be convinced by lone arguments in topics I'm not an expert in, even though there are a lot of really shiny arguments you can adopt to get an edge over the sheeple and strut around knowing The Truth.

> *This is the correct Bayesian action: if I know that a false argument sounds just as convincing as a true argument, argument convincingness provides no evidence either way. I should ignore it and stick with my prior.*

gasp confirmation bias!?

Surely you meant to say determine what the orthodox argument is, and go with that. Hopefully that was your prior, already. Orthodoxy is good. You even describe how following orthodoxy is the only sane way to practice medicine.

Indeed, unless your prior has a detailed comment on why orthodoxy is wrong, you're basically choosing to believe whoever presented an argument on the topic first. Which is how people believe in chemtrails.

HideHide

nobody.really *says:*

June 4, 2019 at 3:16 pm

C.S. Lewis seemed to think that epistemic learned helplessness is a product of modernity.

It sounds as if you supposed that argument was the way to keep him out of the Enemy's clutches. That might have been so if he had lived a few centuries earlier. At that time the humans still knew pretty well when a thing was proved and when it was not; and if it was proved they really believed it. They still connected thinking with doing and were prepared to alter their way of life as the result of a chain of reasoning. But what with the weekly press and other such weapons we have largely altered that. Your man has been accustomed, ever since he was a boy, to have a dozen incompatible philosophies dancing about together inside his head. He doesn't think of doctrines as primarily "true" or "false", but as "academic" or "practical", "outworn" or "contemporary", "conventional" or "ruthless"....

C.S. Lewis, *Screwtape Letters,* (1942), Letter 1.

HideHide

SEE *says:*

June 5, 2019 at 1:28 pm

Yeah, this is a common mistake. For example, you can see Eliezer Yudkowsky making it when comparing the current era to the one when C.S. Lewis was a writer.

The core of the mistake is taking the best thinkers of the past — the ones who authored the works of lasting value — as representative of the past. They aren't.

Hide ↑Hide ↑

nobody.really *says:*

June 4, 2019 at 3:28 pm

You could consider this a form of epistemic learned helplessness, where I know any attempt to evaluate the arguments is just going to be a bad idea so I don't even try. If you have a good argument that the Early Bronze Age worked completely differently from the way mainstream historians believe, I just don't want to hear about it. If you insist on telling me anyway, I will nod, say that your argument makes complete sense, and then totally refuse to change my mind or admit even the slightest possibility that you might be right.

(This is the correct Bayesian action: if I know that a false argument sounds just as convincing as a true argument, argument convincingness provides no evidence either way. I should ignore it and stick with my prior.)

I've pondered precisely this idea as I consider how to play the *Wait Wait Don't Tell Me* "Bluff the Listener" game. Three people tell stories from the week's news, and I need to identify the true one. I can get clues from the stories about how plausible they are.

But then there's Roxanne Roberts: Her stories ALWAYS seem the most plausible–regardless of their accuracy.

So what's the Bayesian strategy? I know that each speaker has a 2/3 likelihood of lying to me, and this applies to Roberts's story, too. And I know that listening to Roberts's story will provide no additional information. So I reject her story out of hand and focus solely on the stories that MIGHT provide me with useful information. Alas, usually those stories sound ludicrous—which, through process of elimination, leads me back to Roberts's story. I guess.

Help!

HideHide

philwelch *says:*

June 4, 2019 at 4:49 pm

You have, perhaps unwittingly, rediscovered [GE Moore's refutation of philosophical skepticism](#).

The general skeptical argument is:

* If you don't know whether or not you're a brain in a vat, you don't know that you have hands.
* You don't know whether or not you're a brain in a vat.
* Therefore, you don't know that you have hands. (Mind blown!)

Moore turns it around:

* (same)
* But I do know that I have hands.
* Therefore I can safely surmise that I am not a brain in a vat.

In general, the Moore technique turns a modus ponens into a modus tollens argument and uses our normal everyday certainty about normal facts to discredit seemingly-plausible logical arguments that tend to refute them. Moore used this to try and dismiss skepticism (in the context of epistemology, meaning the attempt to prove that knowledge is impossible) but the

same technique can be applied against crackpottery, as you demonstrate. I suspect you could construct a Bayesian form of this as well, where you have such a strong prior that eg the moon landing actually happened, you would naturally conclude rather summarily that there existed flaws in the moon-landing-deniers' logic.

Of course, every programmer does this on a daily basis: debugging is by nature the discovery of some subtle, unforeseen flaw in a logical structure that is obviously producing the incorrect result.

HideHide

HeelBearCub *says:*

June 4, 2019 at 7:42 pm

This is a quack argument.

It simply assumes the conclusion.

Hide ↑Hide ↑

philwelch *says:*

June 4, 2019 at 9:34 pm

Not at all. If I write some code and it behaves unexpectedly, should I assume that my code is correct and that I had unfair or incorrect expectations about the behavior? Or should I assume that my code is incorrect?

The only thing that logic can hope to do is to demonstrate an inconsistency. You get to choose how to resolve it, based on which of the inconsistent propositions you would prefer to keep.

Hide ↑Hide ↑

Aapje *says:*

June 4, 2019 at 11:09 pm

> *should I assume that my code is correct and that I had unfair or incorrect expectations about the behavior? Or should I assume that my code is incorrect?*

Code that doesn't meet expectations is incorrect code, so your

example is nonsensical. Code can't be correct and have a mismatch with expectations.

Hide ↑Hide ↑

philwelch *says:*

June 5, 2019 at 1:17 pm

Right, but if I've read the code and found no obvious flaw or inconsistency, I'm at a bit of an impasse. What Scott's article, the Moore reversal, and software debugging all have in common is the act of assuming the existence of a flaw in reasoning that one has yet to discover and pinpoint over accepting a patently absurd conclusion.

If you read a seemingly airtight argument that the years of 614-911 AD never happened, you'd be a fool to prefer the assumption that the argument itself is sound over the assumption that three centuries of history were fabricated. Likewise if you read a seemingly airtight argument that you don't know whether or not you have hands, or if you have seemingly airtight code that does something weird and undesirable.

Hide ↑Hide ↑

HeelBearCub *says:*

June 5, 2019 at 4:13 am

The whole original argument is silly.

There aren't any inconsistencies in steps 1-3 of the original argument. You then invent one for step 4 by assuming your conclusion, that you can know about your physical existence.

Boiled down the entire argument is:

* You don't know whether or not you're a brain in a vat.

* But you do know that you aren't a brain in a vat.

You haven't done anything.

philwelch *says:*

June 5, 2019 at 1:09 pm

The inconsistency is between the propositions, "I know that I am not a brain in a vat" and "I know that I have hands".

HeelBearCub *says:*

June 5, 2019 at 1:23 pm

You munged that.

I assume you mean the inconsistency is between:
* You don't know whether or not you're a brain in a vat.
* You do know that you have hands.

But that simply misunderstands or ignores the first statement. You don't know that you have hands, you only know that you experience the having of hands. If you were a brain in a vat, with fake experiences being pumped in, you would have the same experience.

Thus saying that you "know" you have hands is equivalent to saying you "know" you aren't a brain in a vat.

kalimac *says:*

June 4, 2019 at 7:14 pm

What I've occasionally found is people who will state beforehand what features an argument would need to have in order to convince them. Then I find an argument which has those features, and they start inventing previously unmentioned reasons why it doesn't count.

One specific example I remember – this was from around 30 years ago, when John Lott was still running around unrefuted – was someone who dismissed all gun-control arguments as

specious, and who said he'd take them seriously when he found one in a peer-reviewed scientific journal. Just then an article supporting gun control appeared in the *New England Journal of Medicine*, so I handed to him. He replied that the topic was outside of a medical journal's area of competence, so he rejected it, despite the fact that the article was written in the discipline of public health.

HideHide

Jiro *says:*

June 5, 2019 at 11:02 am

> What I've occasionally found is people who will state beforehand what features an argument would need to have in order to convince them.

The problem isn't that they won't change their mind, the problem is that they shouldn't have made a statement like that.

It's really difficult to describe what would change your mind in a way that's free from loopholes. If you take a well-supported position and you describe what would change your mind about it, most examples you'll get would be people finding loopholes, not people saying things that would substantially refute your position. And the prize of your opponent finding a loophole is that he gets to make you look like a fool, which is a very valuable prize to someone arguing with you and encourages such loophole-finding.

Also, gun control *is* outside a medical journal's area of competence. Public health professionals know the physical effects of gunshots (and of attacks by criminals), but the *relationship* between gun control, gunshots, and criminals is not something they are an expert on. It's like medical professionals saying that we should go to war with China because the war would have a positive economic effect, and in a better economy, people are healthier. The disputed part is the relationship between the war and the economy (and the side effects of war)–not the relationship between the economy and health, which is the only part they might have expertise in.

Hide ↑Hide ↑

Matt M *says:*

June 5, 2019 at 11:27 am

> *It's really difficult to describe what would change your mind in a*

> way that's free from loopholes.

Agreed.

People often ask me things like "What would it take to change your mind?" I either refuse to answer, or flatly state "Nothing can change my mind." Because it is impossible to know, in advance, what sort of evidence I might find persuasive, and why.

Hide ↑Hide ↑

Pattern *says:*

June 5, 2019 at 6:21 am

> But I'm also glad epistemic learned helplessness exists. It seems like a pretty useful social safety valve most of the time.

This seems to me, to be a good argument for skepticism, instead.

HideHide

DawnPaladin *says:*

June 5, 2019 at 8:44 am

> And so they take the obvious and correct defensive maneuver – they will never let anyone convince them of any belief that sounds "weird".

This is why convincing people of climate change is hard. You have a someone who's spent their entire life perfecting their ability to run a business. Then some experts come along from some "climatology" field you've never heard of and insist that you have to change a bunch of stuff or the world's going to end.

HideHide

Matt M *says:*

June 5, 2019 at 9:56 am

Very few actual climatologists are arguing anything close to "the world's going to end."

It is mostly politicians and journalists doing that.

Hide ↑Hide ↑

nobody.really *says:*

June 6, 2019 at 2:48 pm

> *Very few actual climatologists are arguing anything close to "the world's going to end."*
>
> *It is mostly politicians and journalists doing that.*

And cosmologists. And physicists. But they argue that the world is going to end regardless of climate change.

Hide ↑Hide ↑

ariel *says:*

June 5, 2019 at 9:34 am

> *Even the smartest people I know have a commendable tendency not to take certain ideas seriously. Bostrom's simulation argument, the anthropic doomsday argument, Pascal's Mugging – I've never heard anyone give a coherent argument against any of these, but I've also never met anyone who fully accepts them and lives life according to their implications.*

The statistician Olle Haggstrom gives a pretty convincing take down of the Doomsday Argument in his book Here Be Dragons.

I would be perfectly happy to act on the simulation hypothesis if there were any clear implications. I think it's rational to invest at least a token effort in Pascal's wager e.g. von Neumann converting to Catholicism on his deathbed. In general it's not clear which infinities to optimize for, and it makes sense to live a relatively normal life accruing resources and whatnot while your society figures out world-shattering things like how to test the simulation hypothesis or approximate the relative validity of different religions.

HideHide
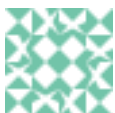
zby *says:*

June 5, 2019 at 1:11 pm

"Inadequate Equilibria" has already been mentioned here – but I'd like second it. It is a whole booklet on the subject of when you should try to trust your arguments and when you really

should just go with the mainstream. It has lots of good critique of how things work in the society now and also ideas on how it could be systematically solved that emit that crackpot feeling. My approach to these solutions is exactly your 'learned helplessness'.

But what should be especially interesting to you is that there is also a medical theme that is kind of leitmotiv of the book – it is about intravenous feeding for short bowel syndrome babies. Apparently those babes are/were at the time when the book was written/ fed in a way that was killing 30% of them when it was already evident that it is easy to fix the formula for years, it just was not mainstreamed because the research articles were somehow of the wrong kind. Somehow in Europe it got mainstreamed but not in the US.

https://equilibriabook.com/molochs-toolbox/ – is the most relevant chapter for the medical theme

HideHide

uau *says:*

June 5, 2019 at 5:40 pm

About the "doomsday argument":

First, the kind of calculations used in the argument can be correct in some contexts. Assume a scenario where someone selects a uniformly random number N between 1 and 1000000. He picks you and N-1 other people, then gives serial numbers in random order to everyone in your group. You get serial 100. What can you tell about N?

Obviously, N can't be less than 100. But serial 100 is also evidence in favor for smaller numbers above 100. Where before learning the serial every number was equally likely, now numbers below 100 are ruled out, and numbers >= 100 are weighted according to 1/n. The probability of N being below 10000 is about 50%. (The serial being 100 is evidence against large N, because with N=100 there's a 1 in 100 chance of getting that serial, but with N=1000000 there's only a 1 in 1000000 chance).

However, I don't think that this logic is applicable to the number of humans, and the doomsday argument is consequently wrong. The argument depends on picking a civilization size randomly *from equally weighted civilizations*, then picking a random person from there. But that's not reasonable. If you imagine a parallel universe for each possible civilization, and you're told you'll be born as a random person in the multiverse, you should expect to be born in one of the larger civilizations. This cancels the knowledge from knowing your serial number. In the original example of someone selecting N between 1 and 1000000 *and guaranteeing you're in the set no*

matter what the size, suppose instead that he selects random N people in your country. You're more likely to be chosen in the cases where N is large. Now, if you get chosen and your serial is 100, each N between 100 and 1000000 is equally likely. To make this more directly parallel the multiverse case, suppose the person running this keeps selecting random N and picking some new N people not selected before until everyone in the country is selected once. This still keeps the "equally likely" property.

I looked at the Wikipedia page but it does not seem to directly contain the above argument. The "Attempted Refutations" part at the end looks like it could refer to the same thing, but it refers to papers in French without clearly saying what those papers contain, and I didn't try reading them.

Edit: to highlight the problem in the "doomsday argument", if you assume a probability distribution where 50% of civilizations end at 10 people and 50% at 100 billion, then the argument does calculations with the assumption that a random person has an a priori equal chance of being in either civilization – that 50% of people are in a civilization of 10 and 50% in a civilization of 100 billion. That's obviously wrong. If *civilizations* have a 50%/50% chance between being either size, then a *person* has a 99.99999999% chance of being in a civilization of 100 billion.

HideHide

---

wanda_tinasky *says:*

June 6, 2019 at 7:14 pm

First of all, your reasoning is wrong. The question can be reframed as: pick a ball from a vat of N balls numbered 1 to N. That's a random sample from the population and can therefore be used to update your estimate of N.

However, your ultimate conclusion is right because being born isn't a random sampling from a preexisting population. We're not a random sampling from the future timeline of all humans, which means our birth order doesn't transmit any information about the number of future humans.

Consider an analogy with a gumball machine: estimate the number of gumballs in a machine by observing the gumballs which come out of the dispenser. (The correspondence to DA should be obvious: gumballs = humans, dispensing a gumball = being born, capacity of the machine = total humans who will ever live.) The argument

people THINK they're making about DA is that they imagine taking a random gumball, marking it, and then seeing what its dispensing order is. That *would* yield a valid estimate of the machine capacity, because it's using a random sample. Unfortunately, humans aren't random samples from a future timeline (mostly because information doesn't travel backwards in time): we simply come into the world without carrying any information about where we came from. All we can observe is our birth order, which gives us information about: our birth order. The gumball analogy is: numbering the balls as they come out of the machine. Which, as I certainly hope is obvious, does not yield any information about the capacity of the machine.

Hide ↑Hide ↑

uau *says:*

June 6, 2019 at 10:33 pm

> *First of all, your reasoning is wrong.*

Which reasoning? You say you agree with the conclusion and do not actually identify any problematic parts...

If you mean that any reasoning based on birth order is wrong, I disagree. Suppose that 50% of civilizations die by the time reach 100 billion humans, and 50% fill the stars. If you somehow appear as a human without any knowledge of which civilization you're from, it's reasonable to assume you're much more likely to be in the civilization with an astronomical number of humans. Learning that you're human number 20 billion is then evidence against being in the large civilization, bringing it back to 50/50. The doomsday argument is correct in saying that birth order is evidence against being in a large civilization – but only in the context of not yet being aware of your civilization, and the argument incorrectly uses this evidence to update the wrong prior probability.

Hide ↑Hide ↑

Ratheka *says:*

June 9, 2019 at 10:06 pm

It's not clear to me what the implications of the simulation argument should lead me to do. I have no way of judging whether I am or am not in sim, but if the numbers are as described, they are not in my favor. Certainly I want to run some ancestor simulations, because there are

people I've lost who I'd like to recover. I accept that I might myself be an ancestor sim someone else is running for this, and… now what? Nothing seems to follow.

HideHide