Comment Permalink

Daniel Kokotajlo 1d 10

Old saying: "He who would sacrifice essential liberty to achieve security deserves neither and will lose both."

I think something like this is typically true with rationality and effective altruism:

"They who imbibe falsehoods in order to win, eventually lose."

It isn't always true, of course. You give some counterexamples above. But it's true often enough, I think, to be worth making into a litany.

Bob Jacobs 1d 1

I think this is obvious, but we shouldn't be afraid to examen when and why our rules fail.

See in context

[Draft] Thought experiments for Epistemic Rationalist

2

by **Bob Jacobs** 9th Jun 2020 15 comments •••

Edit: Moved it to draft by accident, had to re-upload it didn't change anyth

My poll that asked people to choose between either epistemic rationality (aka 'truth') and instrumental rationality (aka 'winning') seems to indicate that about a third of LessWrongers would choose 'truth' over 'winning'. For those people, the people that think truth and winning always go together and the people who are still unsure; I give you 6 thought experiments:

1. Anxi has very severe anxiety, but she is also mildly interested in AI. Right now she's waiting in a waiting room but she has a copy of Roko's Basilisk printed out and ready to read. Because she has nothing else to do anyway she is contemplating whether or not she should read it. She wants to expand her map of the territory (epistemic rationality) but she doesn't want to have panic attacks for the rest of her life (instrumental rationality). What should she do?

2. Budgo is a budget planner for the government. Right now they're deciding whether they should invest twenty million dollars into researching the development of bio-weapons or keep the money for a rainy day. The problem is he suspects that a doomsday cult hellbent on destroying humanity has infiltrated the bio-weapons research group. On the other hand his country rarely invests in intellectual projects, so if he chooses the rainy day option it probably won't go to research. Should Budgo vote to give the money to the research group (epistemically rational) or keep it for a rainy day (instrumentally rational)?

3. Cansa is visting her friend who is fighting with cancer. She hears the doctor tell her friend that he has a 51% chance of being cured. When Cansa confront the doctor that it is actually only a 49% chance the doctor tells her that he knows that. The reason the doctor lied is that studies have shown that people who are told they have a 51% chance of getting

cured get such a psychological boost from that that they then have a 50% chance of being cured. Meanwhile the people who are told that they have a 49% or 50% chance of making it only have a 49% chance of being cured. Unfortunately studies have shown that in this case the deception is necessary, and the placebo effect won't take hold without it. Should Cansa tell her friend?

4. Hazzar gets a mail from his friend. It says that in the attached folder is an information hazard that has been making the rounds on the internet. 90% of the people who have opened it get severe psychological damage and are telling people not to look at it. The remaining 10% say they only learned some mildly interesting but hard to convey information about seagulls. Hazzar is going to open the folder, but should he?

5. Vaco is a scientist that just discovered how you can make a vacuum decay with household objects. While writing his paper he realized that there are no other insights that can come from this knowledge and the process to do it is so strange that there is basically no chance that someone could do it by accident. Now the question remains whether he should publish his results and make this dangerous knowledge available to every lunatic on the planet.

6. You are in a featureless white room. In the center is a big red button with the label "Click here to be tortured". Do you click the button and learn more about this mysterious room or do you not press the button and avoid potentially being tortured?

I won't be criticizing the answers, but please keep in mind that finding a loophole in the question is not the same as solving the question in the least convenient possible world°.

Rationality 2 Practical 2 + Add Tag

2

15 comments, sorted by top scoring Highlighting new comments since Today at 1:18 PM

[-] **shminux** 20h 21

Let me suggest a simpler "thought experiment":

There is a gun loaded as per Russian roulette rules, pointed at your head. Should you pull the trigger to find out "the truth" about its chamber?

My contention is that your other examples are no different.

[-] **jimmy** 16h 13

1) Isomorphic to my "what if you know you'll do something stupid if you learn that your girlfriend has cheated on you" example. To reiterate, any negative effects of learning are caused by false beliefs. Prioritize over which way you're going to be wrong until you become strong enough to just not be predictably wrong, sure. But become stronger so that you can handle the truths you may encounter.

2) This clearly isn't a conflict between epistemic and instrumental rationality. This is a question about arming your enemies vs not doing so, and the answer there is obvious. To reiterate what I said last time, this stuff all falls apart once you realize that these are two entirely separate systems both with their own beliefs and values and you posit that the subsystem in control is not the subsystem that is correct and shares your values. Epistemic rationality doesn't mean giving your stalker your new address.

3) "Unfortunately studies have shown that in this case the deception is necessary, and the placebo effect won't take hold without it". This is assuming your conclusion. It's like saying "Unfortunately, in my made up hypothetical that doesn't actually exist, studies have shown that some bachelors *are* married, so *now* what do you say when you meet a married bachelor!". I say you're making stuff up and that no such thing exists. Show me the studies, and I'll show you where they went wrong.

You can't just throw a blanket over a box and say "now that you can no longer see the gears, imagine that there's a perpetual motion machine in there!" and expect it to have any real world significance. If someone showed me a black box that put out more energy than went into it and persisted longer than known energy storage/conversion mechanisms could do, I would first look under the box for any shenanigans that a magician might try to pull. Next I would measure the electromagnetic energy in the room and check for wireless power transfer. Even if I found none of those, I would first expect that this guy is a better magician than I am anti-magician, and would not begin to doubt the physics. Even if I became assured that it wasn't magician trickery and it really wasn't sneaking energy in somehow, I would then start to suspect that he managed to build a nuclear reactor smaller than I thought possible, or otherwise discovered new physics that makes this possible. I would then proceed to tear the box apart and find out what assumptions I'm missing. At the point where it became likely that it wasn't *new* physics but rather incorrect old physics, I would continually reference the underlying justifications of the laws of thermodynamics and see if I could start to see how one of the founding assumptions could be failing to hold.

Not until I had done all that would I even *start* to believe that it is genuinely what it claims to be. The reasons to believe in the laws of thermodynamics are simply so much stronger than the reason to believe people claiming to have perpetual motion machines that if your first response isn't to challenge the hypothetical *hard*, then you're making a mistake.

"Knowing more true things without knowing more false things leads to worse results by the values of the system that is making the decision even when the system is working properly" is a similarly extraordinary claim that calls for extraordinary evidence. The first thing to look for, besides a complete failure to even meet the description, is for false beliefs being smuggled in. In every case you've given, it's been one or the other of these, and that's not likely to change.

If you want to challenge one of the fundamental laws of rationality, you have to produce a working prototype, and it has to be able to show where the founding assumptions went wrong. You can't simply cast a blanket over the box and declare that it is now "possible" since you "can't see" that is not impossible. Endeavor to open black boxes and see the gears, not close your eyes to them and deliberately reason out of ignorance. Because when you do, you'll start to see the path towards making both your epistemic and your instrumental rationality work better.

4) Throw it away like all spam. Your attention is precious, and you should spend it learning the things that you expect to help you the most, not about seagulls. If you want though, you can use this as an exercise in becoming more resilient and/or about learning about the nature of human psychological frailty.

It's worth noticing though, that you didn't use a real world example and that there might be reasons for this.

5) This is just 2 again.

6) *Maybe?* As stated, probably not. There are a few different possibilities here though, and I think it makes more sense to address them individually.

a) The torture is physically damaging, like peeling ones skin back of slowly breaking every bone in ones body.

In this case, obviously not. I'm also curious what it feels like to be shot in the leg, but the price of that information is more than I'm willing to spend. If I learn what that feels like, then I don't get to learn what I would have been able to accomplish if I could still walk well. There's no conflict here between epistemic and instrumental rationality here.

b) The "torture" is guaranteed to be both safe and non physically damaging, and not keep me prisoner too long when I could be doing other things.

When I learned about tarantula hawks and that their sting was supposedly both debilitatingly painful and also perfectly non-damaging and safe, I went pretty far out of my way to acquire them and provoke them to sting me. Fear of non-damaging things is a failing to be stamped out. When you accept that the scary thing truly is sufficiently non-dangerous, fear just becomes excitement anyway.

If these mysterious white room people think they can bring me a challenge while keeping things sufficiently safe and non-physically-damaging I'd *probably* call their bluff and push that button to see what they got.

c) This "torture" really is enough to push me sufficiently past my limits of composure that there will be lasting psychological damage.

I think this is actually harder than you think unless you also cross the lines on physical damage, risk, or get to spend a *lot* of time at it. However, it is conceivable and so in this case we're back to being another example of number one. If I'm pretty sure it won't be any worse than this, I'd go for it.

This whole "epistemic vs instrumental rationality" thing really is just a failure to do epistemic rationality right, and when you peak into the black box instead of intentionally keeping it covered you can start to see why.

[–] **lsusr** 14h 3

I think your comparison to spam in #4 works well. Reading spam has negative expected utility and small possible positive utility. Negative-sum advertising in general and spam in particular*is* a real-world example, at least in principle.

[–] **Daniel Kokotajlo** 1d 10

Old saying: "He who would sacrifice essential liberty to achieve security deserves neither and will lose both."

I think something like this is typically true with rationality and effective altruism:

"They who imbibe falsehoods in order to win, eventually lose."

It isn't always true, of course. You give some counterexamples above. But it's true often enough, I think, to be worth making into a litany.

[–] **Bob Jacobs** 1d 1

I think this is obvious, but we shouldn't be afraid to examen when and why our rules fail.

[–] **ChristianKl** 20h 6

I wouldn't expect that a very nasty infohazard is likely to increase my epistemic alignment with the world. It could for example be a very convincing deep fake that's engineered in a way to be particularly sticky in my mind.

Part of caring about epistemtic rationality is engaging with sources of information that you think are likely to improve your alignment with reality and not decrease it.

Most of the crowd who believes in epistemic rationality also doesn't see experiential knowledge as being able to provide knowledge in a good way. If a new infohazard has such effects it might be a lot better to read a research article about it then exposing oneselves directly to the piece.

[–] **Dagon** 15h 5

I'm not sure I get it (nor the previous post). Can you steelman (or point out a comment that shows) the point of view you're arguing against? Even for a perfect agent with well-defined goals, it seems obvious that true knowledge which does not approach completeness (that is, there are still lots more unknown things about the universe) can lead to severely sub-optimal choices and future experiences. It seems even clearer that perverse subsets of the truth can be found which will lead an otherwise-rational agent to incorrect action.
For such imperfect agents as we are, with such small access to truth, I just don't understand the debate.

Your examples don't help me understand, as it's not clear what updates the participants are making toward true beliefs. Specifically:

1 - There's no truth value in the description of Roko's Basilisk. I'd argue that learning some actual truth could benefit her (about whether such a thing exists in what percentage of her future timelines), could benefit her.

2 - Wow. Research does not always lead to truth, especially if you don't trust your researchers. Seems likely that saving for future use is best for all purposes. Spending some of it to investigate your research team might be rational, though.

3 - this is a good example, and can help narrow down the claim, if there is a claim to narrow down.

4 - I'd almost certainly open it. I don't think this generalizes to all humans, but I'm pretty smart and I think I'd benefit by understanding a little more about "psychological damage" and infohazard mechanisms.

5 - Vaco's beliefs about his discovery are outlandish enough that he's simply wrong to be confident. Demonstrating it to a respected physicist and delegating the decision to someone more sane seems the obvious answer.

6 - I think there are a lot of ways to get information that I'd try before pressing the button, but I'd likely press it before I die of thirst.

[–] **Isusr** 14h 3

I want to express schadenfreude that Dagon would open the folder in #4. I would also like to note that many people have pressed the button in #6 out of sheer boredom.

[–] **Dagon** 12h 4

I should probably admit that my planned behavior for #4 is mostly hubris, and as such I forgive you your shadenfreude. I deserve what I get (including the sweet sweet knowledge, of course).

But this hubris _is_ based on additional assumptions and knowledge about the universe, including a fair bit about infohazards and "psychological damage", which make me believe the threat is much less than stated. This is a different aspect of the problematic thesis under consideration: these examples are incredibly lacking in information that would allow one to judge the quantity of truth achievable for what amount of risk.

[–] **Bob Jacobs** 14h 1

Thanks you for writing this, yeah it appears that we are talking past each other (particularly in my debate with Jimmy). I was going to write another post to try and clarify this debate, but decided against it for five reasons:

1. I already promised Isusr I would make a different (and hard to make) post this week and I don't want to spam out posts.

2. The downvotes are telling me this is clearly a controversial/community-splitting topic and I've been making too much of those lately

3. I've now already made two posts on this topic so I'm starting to get sick of it.

4. While my english has improved remarkably these past years it is apparently not yet at the level where I can discuss contemporary philosophy without taking the risk of not effectively communicating with my interlocutor (this might sometimes be the fault of your interlocutor but it has happened two times in two weeks so I'm giving myself at the very least partial blame)

5. I have exams right now, so I should really be studying more for those.

Maybe I'll make one in a couple months if no one else has made a post on it yet.

[–] **noggin-scratcher** 12h 4

I would draw a distinction within epistemic rationality between (A) wanting to know more true things, and (B) wanting to avoid believing false things.

Most of the examples given poke at a tension between learning, or potentially learning, or disseminating for others to learn, something new (so, type A epistemics according to the typology I just pulled from my nethers) versus the risks of instrumental harm incurred in the process.

Only example 3 really addresses type B by asking whether it's better for Cansa's friend to believe a falsehood if that will improve their prognosis. But the distinction between believing 51% and 49% is small enough for that to feel like a small and weak falsehood anyway.

I don't know what your poll respondents had in mind, but if you asked *me* about my preference between epistemic/instrumental rationality, truth vs winning, my first thought would be a rather stronger type B example; I'd be asking myself whether I'd want to believe a fairly large falsehood in exchange for instrumental benefits. Which might be sidestepping part of the intent of the poll question, but also might explain some of the apparent discrepancy.

[–] **Bob Jacobs** 3h 1

Yeah but type B and its many forms like placebo, nocebo, psychosis etc etc are already widely known and documented. I only included it begrudgingly for the sake of completeness and because someone was going to mention it in the comments (not sure why I made the effect-size so small). This post is not titled: ‘steelman of the pragmatist position’, otherwise I would have indeed focused more on the type B for which there are more real world examples. I wanted to think up some fun and strange thought experiments that might push peoples brain in directions they don’t usually go and consider new angles.

[–] **ChristianKl** 20h 4

Given that Anxi has an interest in AI. She will hear from time to time about Roko's Basilisk. If she doesn't know what it's about and hears that it's dangerous to engage with it, that is likely produce a lot of anxiety in her.

A decent way for her to deal with the topic would be to seek a friend who actually understands the subject matter and let that friend explain it to her.

I don't see why she should manevour herself into a situation where she's in a waiting room with a printed out copy of Roko's post.

"Have nothing else to do" also doesn't exist because you can always reflect on existing knowledge.

[–] **ChristianKl** 20h 3

> **6.** You are in a featureless white room. In the center is a big red button with the label "Click here to be tortured". Do you click the button and learn more about this mysterious room or do you not press the button and avoid potentially being tortured?

This example seems artificial but I think there are real life equivalents.

I might walk home at night. I have the choice to go through a dark allay and I have fear of going through the alley. Should I go through the dark alley to learn information about the dark alley?

I don't think anybody is going to argue that walking through the alley is an effective way to learn information about what happens in dark alleys.

If you are interesting in learning about dangers of dark alleys you rather read statistics and go through other motions. That's true even if you care more about epistemic rationality then winning.

[–] **ChristianKl** 20h 2

Examples 3 and 5 both assume that we live in a world where it's easy to have highly certain knowledge about very specific questions.

If we would live in such a world, good epistemics wouldn't be as valuable as the world in which we live where it's hard to know things.