

BEST OF LESSWRONG 2018

# Local Validity as a Key to Sanity and Civilization

by Eliezer Yudkowsky 7th Apr 2018

Review by Ben Pace

Review by Zack\_M\_Davis

Review by Zvi

Review by orthonormal

Review by Jameson Quinn

(Cross-posted [from Facebook](#).)

0.

Tl;dr: There's a similarity between these three concepts:

- A locally valid proof step in mathematics is one that, in general, produces only true statements from true statements. This is a property of a single step, irrespective of whether the final conclusion is true or false.
- There's such a thing as a bad argument even for a good conclusion. In order to arrive at sane answers to questions of fact and policy, we need to be curious about whether arguments are good or bad, independently of their conclusions. The rules against fallacies must be enforced even against arguments for conclusions we like.
- For civilization to hold together, we need to make coordinated steps away from Nash equilibria in lockstep. This requires general rules that are allowed to impose penalties on people we like or reward people we don't like. When people stop believing the general rules are being evaluated sufficiently fairly, they go back to the Nash equilibrium and civilization falls.

## i.

The notion of a locally evaluated argument step is simplest in mathematics, where it is a formalizable idea **in model theory**<sup>o</sup>. In math, a general type of step is 'valid' if it only produces semantically true statements from other semantically true statements, relative to a given model. If  $x = y$  in some set of variable assignments, then  $2x = 2y$  in the same model. Maybe  $x$  doesn't equal  $y$ , in some model, but even if it doesn't, the local step from " $x = y$ " to " $2x = 2y$ " is a locally valid step of argument. It won't introduce any *new* problems.

Conversely,  $xy = xz$  does not imply  $y = z$ . It happens to work when  $x = 2$ ,  $y = 3$ , and  $z = 3$ , in which case the two statements say " $6 = 6$ " and " $3 = 3$ " respectively. But if  $x = 0$ ,  $y = 4$ ,  $z = 17$ , then we have " $0 = 0$ " on one side and " $4 = 17$ " on the other. We can feed in a true statement and get a false statement out the other end. This argument is not *locally* okay.

You can't get the concept of a "mathematical proof" unless on some level—though often an intuitive level rather than an explicit one—you understand the notion of a single step of argument that is locally okay or locally not okay, independent of whether you globally agreed with the final conclusion. There's a kind of approval you give to the pieces of the argument, rather than looking the whole thing over and deciding whether you like what came out the other end.

Once you've grasped that, it may even be possible to convince you of mathematical results that sound counterintuitive. When your understanding of the rules governing allowable argument steps has become stronger than your faith in your ability to judge whole intuitive conclusions, you may be convinced of truths you would not otherwise have grasped.

## ii.

More generally in life, even outside of mathematics, there are such things as bad arguments for good conclusions.

There are even such things as genuinely good arguments for false conclusions, though of course those are much rarer. By the Bayesian definition of evidence, "strong evidence" is exactly that kind of evidence which we very rarely expect to find supporting a false conclusion. Lord Kelvin's careful and multiply-supported lines of reasoning arguing that the Earth could not possibly be so much as a hundred million years old, all failed simultaneously in a surprising way because

that era didn't know about nuclear reactions. But most of the time this does not happen.

On the other hand, bad arguments for true conclusions are extremely easy to come by, because there are tiny elves that whisper them to people. There isn't anything the least bit more difficult in making an argument terrible when it leads to a good conclusion, since the tiny elves own lawnmowers.

One of the marks of an intellectually strong mind is that they are able to take a curious interest in whether a particular argument is a good argument or a bad argument, independently of whether they agree with the conclusion of that argument.

Even if they happen to start out believing that, say, the intelligence explosion thesis for Artificial General Intelligence is false, they are capable of frowning at the argument that the intelligence explosion is impossible because hypercomputation is impossible, or that there's really no such thing as intelligence **because of the no-free-lunch theorem**, and saying, "Even if I agree with your conclusion, I think that's a terrible argument for it." Even if they agree with the mainstream scientific consensus on anthropogenic global warming, they still wince and perhaps even offer a correction when somebody offers as evidence favoring global warming that there was a really scorching day last summer.

There are weaker and stronger versions of this attribute. Some people will think to themselves, "Well, it's important to use only valid arguments... but there was a sustained pattern of record highs worldwide over multiple years which does count as evidence, and that particular very hot day was a part of that pattern, so it's valid evidence for global warming." Other people will think to themselves, "I'd roll my eyes at someone who offers a single very cold day as an argument that global warming is false. So it can't be okay to use a single very hot day to argue that global warming is true."

I'd much rather buy a used car from the second person than the first person. I think I'd pay at least a 5% price premium.

Metaphorically speaking, the first person will court-martial an allied argument if they must, but they will favor allied soldiers when they can. They still have a sense of motion toward the Right Final Answer as being progress, and motion away from the right final answer as anti-progress, and they dislike not making progress.

The second person has something more like the strict mindset of a mathematician when it comes to local validity. They are able to praise some proof steps as obeying

when it comes to local validity. They are able to praise some proof steps as obeying the rules, irrespective of which side those steps are on, without a sense that they are thereby betraying their side.

### iii.

This essay has been bubbling in the back of my mind for a while, since I read that potential juror #70 for the Martin Shkreli trial was rejected during selection when, asked if they thought they could render impartial judgment, they replied, "I can be fair to one side but not the other." And I thought maybe I should write something about why that was possibly a harbinger of the collapse of civilization. I've been musing recently about how a lot of the standard Code of the Light isn't really written down anywhere anyone can find.

The thought recurred during the recent #MeToo saga when some Democrats were debating whether it made sense to kick Al Franken out of the Senate. I don't want to derail into debating Franken's behavior and whether that degree of censure was warranted per se, and I'll delete any such comments. What brought on this essay was that I read some unusually frank concerns from people who did think that Franken's behavior was per se cause to not represent the Democratic Party in the Senate; but who worried that the Democrats would police themselves, the Republicans wouldn't, and so the Republicans would end up controlling the Senate.

I've heard less of that since some upstanding Republican voters in Alabama stayed home on election night and put Doug Jones in the Senate.

But at the time, some people were replying, "That seems horrifyingly cynical and realpolitik. Is the idea here that sexual line-crossing is only bad and worthy of punishment when Republicans do it? Are we deciding that explicitly now?" And others were saying, "Look, the end result of your way of doing things is to just hand over the Senate to the Republican Party."

This is a conceptual knot that, I'm guessing, results from not explicitly distinguishing game theory from goodness.

There is, I think, a certain intuitive idea that *ideally* the Law is supposed to embody a subset of morality insofar as it is ever wise to enforce certain kinds of goodness. Murder is bad, and so there's a law against this bad behavior of murder. There's a lot of places where the law is in fact evil, like the laws criminalizing marijuana; that means the law is departing from its purpose, falling short of what it should be. Those who are not real-life straw authoritarians (who are sadly

should be. Those who are not real-life straw authoritarians (who are sadly common) will cheerfully agree that there are some forms of goodness, even most forms of goodness, that it is not wise to try to legislate. But insofar as it *is* ever wise to make law, there's an intuitive sense that law should reflect some particular subset of morally good behavior that we have decided it is wise to enforce with guns, such as "Don't kill people."

It's from this perspective that "As a matter of pragmatic realpolitik we are going to not enforce sexual line-crossing rules against Democratic senators" seems like giving up, and maybe a harbinger of the fall of civilization if things have really gotten that bad.

But there's more than one function of legal codes, the way that money is both a store of value and a medium of exchange but these are different functions of money.

You can also look at laws as a kind of game theory played with people who might not share your morality at all. Some people take this perspective almost exclusively, at least in their verbal reports. They'll say, "Well, yes, I'd like it if I could walk into your house and take all your stuff, but I would dislike it even more if you could walk into my house and take *my* stuff, and that's why we have laws." I'm never quite sure how seriously to take the claim that they'd be happy walking into my house and taking my stuff. It seems to me that law enforcement and even social enforcement are simply not effective enough to count for the vast majority of human cooperation, and I have a sense that civilization is free-riding a whole lot on innate altruism... but game theory is certainly *a* function served by law.

The same way that money is both medium of exchange and store of value, the law is both collective utility function fragment and game theory.

In its function as game theory, the law (ideally) enables people with different utility functions to move from bad Nash equilibria to better Nash equilibria, closer to the Pareto frontier. Instead of mutual defection getting a payoff of (2, 2), both sides pay 0.1 for law enforcement and move to enforced mutual cooperation at (2.9, 2.9).

From this perspective, everything rests on notions like "fairness", "impartiality", "equality before the law", "it doesn't matter whose ox is being gored". If the so-called law punishes your defection but lets the other's defection pass, and this happens systematically enough and often enough, it is in your interest to blow up the current equilibrium if you have a chance.

It is coherent to say, "Crossing this behavioral line is universally bad when anyone

does it, and also we're not going to punish Democratic senators unless you also punish Republican senators." Though as the saga of Senator Doug Jones of Alabama also shows, you should be careful about preemptively assuming the other side won't cooperate; there are sad lost opportunities there.

#### iv.

The way humans do law, it depends on the existence of what *feel like* simple general rules that apply to all cases.

This is not a universal truth of decision theory, it's a consequence of our cognitive limitations. Two superintelligences could negotiate a compromise with complicated detailed boundaries going right up to the Pareto frontier. They could agree on mutually verified pieces of cognitive code designed to intelligently decide future events according to known principles.

Humans use simpler laws than that.

To be clear, the kind of "law" I'm talking about here is not to be confused with the enormous modern morass of unreadable regulations. Think of, say, the written laws that actually got enforced in a small town in California in 1820. Or Democrats debating whether to enforce a sanction against Democratic senators if it's not being enforced against Republican senators. Or a small community's elders' star-chamber meeting to debate an accusation of sexual assault. Or the laws that cops will enforce even against other cops. These are the kinds of laws that must be simple in order to exist.

The reason that hunter-gatherer tribes don't have 100,000 pages of written legalism... is *not* that they've wisely realized that lengthy rules are easier to fill with loopholes, and that complicated regulations favor large corporations with legal departments, and that laws often have unintended consequences which don't resemble their stated justifications, and that deadweight losses increase quadratically. It's *very clear* that a supermajority of human beings are not that wise. Rather, hunter-gatherers just don't have enough time, energy, and paper to screw up that badly.

When humans try to verbalize The Law that isn't to be confused with written law, the law that cops will enforce against other cops, it comes out in universally quantified short sentences like "Anyone who defects in the Prisoner's Dilemma will be penalized TEN points even if that costs us fifteen" or "If you kill somebody who wasn't attacking you first, we'll exile you."

At one point somebody had the bright idea of trying to write down The Law. That way everyone could have common knowledge of what The Law was; and if you didn't break what was written, you could know you were safe from at least the official sanctions. Robert Heinlein called it the most important moment in political history, declaring that the law was above the politicians.

I for one rather doubt the Code of Hammurabi was universally enforced. I expect that hunter-gatherer tribes long before writing had a sense of there being Laws that were above the decisions of individual elders. I suspect that even in the best of times most of the The Law was never written down, and that more than half of what was written down was never really The Law.

But unfortunately, once somebody had the bright idea of writing down The Law, somebody else had the bright idea of writing down more words on the same clay tablet.

Today we live in a post-legalist era, when almost all of that which serves the true function of Law can no longer be written down. The government legalist system is too expensive in time and money and energy, too unreliable, and too slow, for any sane victim of sexual assault to appeal to the criminal justice system instead of the media justice system or the whispernet justice system. The civil legalist system outside of small claims court is a bludgeoning contest between entities that can afford lawyers, and the real law between corporations is enforced by merchant reputation and the threat of starting a bludgeoning contest. If you're in a lower-class neighborhood in the US, you can't get together and create order using your own town guards, because the police won't allow it. From your perspective, the function of the police is to prevent open gunfights and to not allow any more effective order than that to form.

But so it goes. We can't always keep the nice things we used to have, like written laws. The privilege was abused, and has been revoked.

When remains of The Law must indeed be simple, because our written-law privileges have been revoked, and so The Law relies on everyone knowing The Law without it being written down. It isn't even recited in memorable verse, as once it was. The Law relies on the community agreeing on the application of The Law without there being professional judges or a precedent-based judiciary. If not universal agreement, it must at least seem that the choices of the elders are trying to appeal to The Law instead of just naked self-interest. To the extent a voluntary association can't agree on The Law in this sense, it will soon cease to be a voluntary

. . .

association.

The Law also breaks down if people start believing that, when the simple rules say one thing, the deciders will instead look at whose ox got gored, evaluate their personal interest, and enforce a different conclusion instead.

Which is to say: human law ends up with what people at least *believe* to be a set of simple rules that can be locally checked to test okay behavior. It's not actually algorithmically simple any more than walking is cheaply computable, but it feels simple the way that walking feels easy. Whatever doesn't feel like part of that small simple set won't be systematically enforced by the community, regardless of whether your civilization has reached the stage where police are seizing the cars of black people but not white people who use marijuana.

## V.

The game-theoretic function of law can make following those simple rules feel like losing something, taking a step backward. You don't get to defect in the Prisoner's Dilemma, you don't get that delicious (5, 0) payoff instead of (3, 3). The law may punish one of your allies. You may be losing something according to your actual value function, which *feels like* the law having an objectively bad immoral result. You may coherently hold that the universe is a worse place for an instance of the enforcement of a good law, relative to its counterfactual state if that law could be lifted in just that instance without affecting any other instances. Though this does require seeing that law as having a game-theoretic function as well as a moral function.

So long as the rules are seen as moving from a bad global equilibrium to a global equilibrium seen as better, and so long as the rules are mostly-equally enforced on everyone, people are sometimes able to take a step backward and see that larger picture. Or, in a less abstract way, trade off the reified interest of The Law against their own desires and wishes.

This mental motion goes by names like "justice", "fairness", and "impartiality". It has ancient exemplars like a story I couldn't seem to Google, about a Chinese general who prohibited his troops from looting, and then his son appropriated a straw hat from a peasant; so the general sentenced his own son to death with tears running down his eyes.

Here's a fragment of thought as it was before the Great Stagnation, as depicted in passing in H. Beam Piper's *Little Fuzzy*, one of the earliest books I read as a child.

It's from 1962, when the **memetic collapse** had started but not spread very far into science fiction. It stuck in my mind long ago and became one more tiny little piece of who I am now.

"Pendarvis is going to try the case himself," Emmert said. "I always thought he was a reasonable man, but what's he trying to do now? Cut the Company's throat?"

"He isn't anti-Company. He isn't pro-Company either. He's just pro-law. The law says that a planet with native sapient inhabitants is a Class-IV planet, and has to have a Class-IV colonial government. If Zarathustra is a Class-IV planet, he wants it established, and the proper laws applied. If it's a Class-IV planet, the Zarathustra Company is illegally chartered. It's his job to put a stop to illegality. Frederic Pendarvis' religion is the law, and he is its priest. You never get anywhere by arguing religion with a priest."

There is no suggestion in 1962 that the speakers are gullible, or that Pendarvis is a naif, or that Pendarvis is weird for thinking like this. Pendarvis isn't the defiant hero or even much of a side character. It's just a kind of judge you sometimes run into, part of a normal environment as projected from the author's mind that wrote the story.

If you don't have some people like Pendarvis, and you don't appreciate what they're trying to do even when they rule against you, sooner or later your tribe ends.

I mean, I doubt the United States will literally fall into anarchy this way before the AGI timeline runs out. But the concept applies on a smaller scale than countries. It applies on a smaller scale than communities, to bargains between three people or two.

The notion that you can "be fair to one side but not the other", that what's called "fairness" is a kind of favor you do for people you like, says that even the *instinctive* sense people had of law-as-game-theory is being lost in the modern memetic collapse. People are being exposed to so many social-media-viral depictions of the Other Side defecting, and viewpoints exclusively from Our Side without any leavening of any other viewpoint that might ask for a game-theoretic compromise, that they're losing the ability to appreciate the kind of anecdotes they used to tell in ancient China.

(Or maybe it's hormonelike chemicals leached from plastic food containers. Let's

not forget all the psychological explanations offered for a wave of violence that turned out to be lead poisoning.)

vi.

And to take the point full circle:

The mental motion to evenhandedly apply The Rules irrespective of their conclusion is a kind of thinking that human beings appreciate intuitively, or at least they appreciated it in ancient China and mid-20th-century science fiction. In fact, we appreciate The Law more natively than we appreciate the notion of local syntactic rules capturing semantically valid steps in mathematical proofs, go figure.

So the legal metaphor is where a lot of people get started on epistemology: by seeing the local rules of valid argument as The Law, fallacies as crimes. The unusually healthy of mind will reject bad allied arguments with an emotional sense of practicing the way of an impartial judge.

It's ironic, in a way, because there is no game theory and no morality to the true way of the map that reflects the territory. A paperclip maximizer would also strive to debias its cognitive processes, alone in its sterile universe.

But I would venture a guess and hypothesis that you are better off buying a used car from a random mathematician than a random non-mathematician, even after controlling for IQ. The reasoning being that mathematicians are people whose sense of Law was strong enough to be appropriated for proofs, and that this will correlate, if imperfectly, with mathematicians abiding by what they see as The Law in other places as well. I could be wrong, and would be interested in seeing the results of any study like this if it were ever done. (But no studies on self-reports of criminal behavior, please. Unless there's some reason to believe that the self-report metric isn't measuring "honesty times criminality" rather than "criminality".)

I have no grand agenda in having said all this. I've just sometimes thought of late that it would be nice if more of the extremely basic rules of thinking were written down.

Mentioned in

226 Elements of Rationalist Discourse

218 The Sun is big, but superintelligences will not spare Earth a little sunlight

184 Decoupling vs Contextualizing Norms

160 Biology-Inspired AGI Timelines: The Trick That Never Works

## 150 Meta-Honesty: Firming Up Honesty Around Its Edge-Cases

70 comments, sorted by top scoring

[-] **zulupineapple** 8y ▼ 62 ▲ × 2 ✓

The post repeatedly implies that the situation used to be better, and that it is getting worse, while only providing the weakest possible evidence for the trend. As someone who appreciates the Law, I'm disturbed by that. How do you, in one post, both criticize the argument "*global warming is true because it was very hot yesterday*" and make the argument "*people used to appreciate the Law more because I've read a couple of fictional stories that suggest so*"?

[-] **Rob Bensinger** 8y ▼ 26 ▲ × 3 ✓

Not every conclusion is easy to conclusively demonstrate to arbitrary smart readers with a short amount of text, so I think it's fine for people to share their beliefs without sharing all the evidence that got them there, and it's good for others to flag the places where they disagree and will need to hear more. I think "... because I've read a couple of fictional stories that suggest so" is misunderstanding Eliezer's reasoning / failing his ITT. (Though it's possible what you meant is that he should be *explicit about the fact* that he's relying on evidence/arguments that lack a detailed canonical write-up? That seems reasonable to me.)

[-] **zulupineapple** 8y ▼ 37 ▲ × 2 ✓

This post is hardly a "short amount of text" and this view is quite central to the post, but, sure, I can imagine that EY is hiding some stronger arguments for his view that he didn't bother to share. That's not the problem. The problem is that his post about the value of local validity is not locally valid (in the sense that he himself suggests). The argument, about how people appreciate the Law less now, that he makes, is exactly as valid or invalid as the argument, about global warming, that he criticizes (the validity of both arguments is somewhat debatable).

Some people will think to themselves, "Well, it's important to use only valid arguments... but there was a sustained pattern of record highs worldwide over multiple years which does count as evidence, and that particular very hot day was a part of that pattern, so it's valid evidence for global warming."

Remember this paragraph? I think you might be one of these people. What you said is all technically true, but would you really use the same argument to defend someone in your outgroup? You see, I could take your post, change a few words, and get a comment that criticizes EY for failing to pass the ITT of the climate change denier/supporter (your first sentence is in fact fully general, it can support anything without modifications). It would be a valid and reasonable argument, so I wonder why you didn't make it.

[-] **Rob Bensinger** 8y ▼ 13 ▲ × -1 ✓

This post is hardly a "short amount of text"

I think a satisfactory discussion of the memetic collapse claim would probably have to be a lot longer, and a lot of it would just be talking about more data points and considering different interpretations of them.

I think the criticism "isolated data points can cause people to over-update when they're presented in vivid, concrete terms" makes sense, and this is a big part of why it's pragmatically

valuable to push back against "one hot day ergo climate change", because even though it's nonzero Bayesian evidence for climate change, the strength of evidence is way weaker than the emotional persuasiveness. I don't have a strong view on whether Eliezer should add some more caveats in cases like this to ensure people are aware that he hasn't demonstrated the memetic collapse thesis here, vs. expecting his readers to appropriately discount vivid anecdotes as a matter of course. I can see the appeal of both options.

I think the particular way you phrased your objection, in terms of "is this a locally valid inference?" rather than "is this likely to be emotionally appealing in... (read more)

[] **zulupineapple** 8y 13  0 

The problem is that we don't have the Law for bad arguments written down well enough. You have your ideas of what is bad, I have mine, but I think pointing out irony is still the best most universally acceptable thing I can do. Especially since we're in the comments of a post called "Local Validity as a Key to Sanity and Civilization".

"isolated data points can cause people to over-update when they're presented in vivid, concrete terms"

This is a valid concern, but it is not a valid law. Imagine someone telling you "*your evidence is valid, but you presented it in overly vivid, concrete terms, so I downvoted you*". It would be frustrating. Who decides what is and isn't too emotionally persuasive? Who even measures or compares persuasiveness? That sort of rule is unenforceable.

I think you're making overly strong symmetry claims here in ways that make for a cleaner narrative.

Oh absolutely, there are many differences between the two claims, though my comparison is less charitable for EY than yours. Let  $H$  be "*global warming*" and let  $E$  be "*a random day was hot*". Then  $P(E|H) > P(E|\text{not } H)$  is a mathematically true fa... (read more)

[] **Rob Bensinger** 8y 6  0 

However the correct response is not the take the single data point provided more charitably.

You're conflating two senses of "take a single data point charitably": (a) "treat the data point as relatively strong evidence for a hypothesis", and (b) "treat the author as having a relatively benign reason to cite the data point even though it's weak". The first is obviously bad (since we're assuming the data is weak evidence), but you aren't claiming I did the first thing. The second is more like what I actually... (read more)

[] **zulupineapple** 8y 10  0 

(a) "treat the data point as relatively strong evidence for a hypothesis", <...>. The first is obviously bad (since we're assuming the data is weak evidence), but you aren't claiming I did the first thing.

Honestly, I'm not sure what you did. You said I should distinguish claims that can have short arguments and claims that can't. I assumed that by "distinguish" you meant we

short arguments and claims that can't. I assumed that by "distinguish", you meant we should update on the two claims differently, which sounds like (a). What did "distinguish" really mean?

(b) "treat the author as having a relatively benign reason to cite the data point even though it's weak"

I wasn't considering malicious/motivated authors at all. In my mind the climate supporter either doesn't know about long term measurements, or doesn't trust them for whatever reason. Sure, a malicious author would prefer using weak evidence when strong evidence exists, but they would also prefer topics where strong evidence doesn't exist, so ultimately I don't know in what way I should distinguish the two claims in relation to (b).

A simple counterexample is "I assign 40:1 odds that my friend Bob has personali  
... (read more)

[[-\]](#) namespace 8y 11 0

EY read more than 'a couple of fictional stories'. But I think his pointing toward the general degradation of discourse on the Internet is reasonable. Certainly some segments of Tumblr would seem to be a new low acting as a harbinger of the end times. :P

The problem with this sort of hypothesis, is that it's very hard to prove rigorously. And the reason that's a problem is sometimes hypothesis that are hard to prove rigorously happen to be true anyway. The territory does not relent for a bit because you haven't figured out how to prove your point. **People still get lead poisoning even if the levers of authority insist your argument for toxicity is groundless.** That's a large part of why **I think of measurement as the queen of science.** If you can observe things but aren't entirely sure what to make of the observations, that makes it hard to really do rigorous science with them.

[[-\]](#) zulupineapple 8y 13 0

The person who says that it was hot yesterday also remembers more than one hot day, but that doesn't make their argument much stronger. In fact, even if EY had read all fiction books in the last 100 years, and counted all the Law abiding characters in them by year, that still wouldn't be a strong argument.

the general degradation of discourse on the Internet is reasonable.

He didn't say anything about the internet. I'm pretty sure he's talking about general public discourse. The internet is very new, and mainstream discourse on it is even newer, so drawing trends from it is a bit fishy. And it's not clear that those trends would imply anything at all about general public discourse.

The problem with this sort of hypothesis, is that it's very hard to prove rigorously. And the reason that's a problem is sometimes hypothesis that are hard to prove rigorously happen to be true anyway.

I feel like you're doing something this EY's post is arguing against.

[–] **TheWakalix** 8y 1 0

I feel like you're doing something this EY's post is arguing against.

Care to specify how that is the case?

[–] **zulupineapple** 8y 5 0

I'm suggesting that he (Hypothesis) is making an argument that's almost reasonable, but that he probably wouldn't accept if the same argument was used to defend a statement he didn't agree with (or if the statement was made by someone of lower status than EY).

It might be true that EY's claim is very hard to prove with any rigor, but that is not a reason to accept it. The text of EY's post suggests that he is quite confident in his belief, but if he has no strong arguments (and especially if no strong arguments can exist), then... (read more)

[–] **Rob Bensinger** 8y 10 0

I'm suggesting that he (Hypothesis) is making an argument that's almost reasonable, but that he probably wouldn't accept if the same argument was used to defend a statement he didn't agree with (or if the statement was made by someone of lower status than EY).

This kind of claim is plausible on priors, but I don't think you've provided Bayesian evidence in this case that actually discriminates pathological ingroup deference from healthy garden-variety deference. "You're putting more stock in a claim because you agree with other things the claimant has said" isn't in itself doing epistemics wrong.

In a community where we try to assign status/esteem/respect based on epistemics, there's always some risk that it will be hard to notice evidence of ingroup bias because we'll so often be able to say "I'm not biased; I'm just correctly using evidence about track records to determine whose views to put more weight on". I could see an argument for having more of a presumption of bias in order to correct for the fact that our culture makes it hard to spot particular instances of bias when they do occur. On the other... (read more)

[–] **Rob Bensinger** 8y 9 0

In a community where we try to assign status/esteem/respect based on epistemics, there's always some risk that it will be hard to notice evidence of ingroup bias because we'll so often be able to say "I'm not biased; I'm just correctly using evidence about track records to determine whose views to put more weight on". I could see an argument for having more of a presumption of bias in order to correct for the fact that our culture makes it hard to spot particular instances of bias when they do occur. On the other hand, being

... (read more)

[–] **zulupineapple** 8y 6 0

[Status Regulation and Anxious Underconfidence°](#) and [Hero Licensing°](#).

Would it surprise you to know that I have issues with those posts as well?

[–] **zulupineapple** 8y 1 0

"bias!"

On one hand, I'd much rather talk about how valid "memetic collapse" is, then about how valid someone's response to "memetic collapse" is. On the other hand, I really do believe that the response to this post is a lot less negative than it should be. Then again, these are largely the same question: why is my reaction to this post seemingly so different from other users'? "Bias" isn't necessarily my favorite answer. Maybe they're all just very polite.

"You're putting more stoc

... (read more)

[–] **Rob Bensinger** 8y 13 0

It's not wrong, but it's not locally valid. Here again, I'm going for that sweet irony.

If local validity meant never sharing your confidence levels without providing all your evidence for your beliefs, local validity would be a bad desideratum.

I could trust EY to be right, but personally I don't. Therefore, EY's post didn't really force me to update my estimate of  $P(\text{"memetic collapse"})$  in either direction.

Yes. I think that this is a completely normal state of affairs, and if it *doesn't* happen very often then there's probably something very wrong with the community's health and epistemic hygiene:

- Person A makes a claim they don't have time to back up.
- Person B trusts A's judgment enough to update nontrivially in the direction of the claim. B says as much, but perhaps expresses an interest in hearing the arguments in more detail (e.g., to see if it makes them update further, or out of intellectual curiosity, or to develop a model with more working parts, or to do a spot check on whether they're correct to trust A that much).
- Person C doesn't trust A's (or, implicitly, B's) judgment enough to make a non

... (read more)

[–] **zulupineapple** 8y 7 0

I've said that I think your criticisms have been inconsistent, unclear, or equivocation-prone on a lot of points

Elaborate please. My claims about EY's "memetic collapse" should be clear and simple: it's a bad idea supported by bad arguments. My claims about how reasonable your

it's a bad idea supported by bad arguments. My claims about how reasonable your response to "memetic collapse" is, are much weaker and more complicated. This is largely because I can't read your mind, and you haven't shared your reasoning much. What was your prior for "memetic collapse" before you read this?... (read more)

[] **TheWakalix** 8y 2 -1

The level of charity you are exhibiting is ridiculous. Your arguments are fully general. You could take any post, no matter how stupid, and say "the author didn't have time to share his hard-to-transmit evidence", in defense of it. This is not healthy reasoning.

If Fully General Counterargument A exists, but is invalid, then any defense against Counterargument A will necessarily also be Fully General.

[] **zulupineapple** 8y 4 0

I don't understand what you're trying to say. All fully general arguments are invalid, and pointing out that an argument is fully general is a reasonable defence against it. This defence is not fully general, in the sense that it only works when the original argument is, in fact, fully general.

[] **TheWakalix** 8y 4 0

Rob isn't saying that "complex ideas are hard to quickly explain" supports Yudkowsky's claim. He's saying that it weakens your argument against Yudkowsky's claim. The generality of Rob's argument should be considered relative to what he's defending *against*. You are saying that since the defense can apply to any complex idea, it is fully general. But it's a defense *against* the implied claim that only quick-to-explain ideas are valid.

A fully general counter-argument can attack all claims equally. A good defense ag... (read more)

[] **zulupineapple** 8y 7 0

Here is what our exchange looks like from my point of view.

**Me:** EY's arguments are bad.

**Rob:** But EY didn't have time to transmit his evidence.

Indeed he is not saying "EY is correct". But what *is* he saying? What is the purpose of that reply? In what way is it a reasonable reply to make? I'd love to hear an opinion from you as a third party.

Here is my point of view. I'm trying to evaluate the arguments, and see if I want to update P("memetic collapse") as well as P("EY makes good arguments") or P("EY is a ... (read more)

[] **TheWakalix** 8y 5 0

As I see it, Rob is defending the use of [(possibly shared) intuition?] in an argument. since not everything can be feasibly and quickly proved rigorously to

the satisfaction of everyone involved:

These are the kinds of claims where it's certainly possible to reach a confident conclusion if (as it happens) the effect size is large, but where there will be plenty of finicky details and counter-examples and compressing the evidence into an easy-to-communicate form is a pretty large project. A skeptical interlocutor in those cases could reasonably doubt t

... (read more)

[[-](#)] **zulupineapple** 8y 9 0

Not every post needs to prove the validity of each concept it connects with.

Nobody ever said that it does. It's ok not to give any arguments. It's bad when you do give arguments and those arguments are bad. Can you confirm whether you see any arguments in the OP and whether you find them logically sound? Maybe I am hallucinating.

Yudkowsky simply wasn't trying to convince a skeptic of memetic collapse in this post

That would be fine, I could almost believe that it's ok to give bad arguments when the purpose of the post is different. But t... (read more)

[[-](#)] **FoxLisk** 2y 4 2

I'm a bit late to this but I'm glad to see that you were pointing this stuff out in thread. I see this post as basically containing 2 things:

1. some useful observations about how the law (and The Law) requires even-handed application to serve its purpose, and how thinking about the law at this abstract level has parallels in other sorts of logical thinking such as the sort mathematicians do a lot of. this stuff feels like the heart of the post and i think it's mostly correct. i'm unsure how convinced i would be if i didn't already mostly agree with it, though

... (read more)

[[-](#)] **clone of saturn** 8y 1 0

Eliezer explains his reasoning [here](#), which was linked in the post.

[[-](#)] **zulupineapple** 8y 2 0

Yes, I saw, but I don't find the arguments there more compelling (rather, it's mostly just the same fiction argument). To be clear, the claim that *something* is changing isn't that controversial. It's the claim that this change is "bad", that needs stronger arguments.

[ - ] **Jacob Falkovich** 8y 36 2

This is critically important, and I want to focus on where *our* community can establish better norms for local validity. There are some relevant legal issues (is everyone here planning to pay capital gains tax on their Bitcoins?), but our core concern is epistemic.

We usually do a good job of not falling for outright fallacies, but there's always a temptation to sort of *glide* over them **in the pursuit of a good cause**. I see this a lot in conversations about effective altruism that go something like this:

**Someone:** Cash transfers have an even larger impact than the dollar amount itself, because people mostly use it to buy income-generating assets such as a cow or a sowing machine.

**Me:** Really? I thought most of the money was spent on quality-of-life things like metal roofs and better food.

**Someone:** What, you think that malnourished people don't deserve to buy a bit of food?

This is different from doubling down and claiming that food is actually an income-generating asset because well-fed people make more money. It's a nimble change of subject that prevents the person from actually admitting that their argument was bad.

I don't think it's as big a problem as **some peop**... (read more)

[ - ] **zulupineapple** 8y 8 0

I'd like nothing more than that. Unfortunately, this is hard to do. It's not like you can formalize and then formally verify all arguments. Do you have ideas? I wish there was more discussion on this in LW.

[ - ] **Raemon** 8y 13 0

Formally verifying all arguments takes work, and how easy it is depends on how nuanced or complex a given argument is. But noticing and pointing out obvious errors is something I think most people can do.

In Jacob's exchange above, a passerby can easily notice that the goalposts have shifted and say "hey, it seems like you're shifting your goalposts."

Whether this works depends a lot on whether people expect that sort of criticism to be leveled fairly. In many corners of the internet, people (rightly) assume that complaining about goal-post-moving is a tactic employed against your outgroup.

I'm quite optimistic about solving this problem on LessWrong in particular. I'm much less optimistic about solving it in all corners of the EA-and-Rationalsphere.

[ - ] **zulupineapple** 8y 9 -1

Jacob's example is presumably a caricature, but even then, with a little charity, I can empathize with Someone. Here's how I might translate the debate:

**Someone:** I believe hypothesis "*most people buy cows*" is most likely, therefore "*cash-transfers are good*".

**Jacobian:** Really? I believe hypothesis "*most people buy food*" is more likely (presumably I'm saying this to imply "*cash-transfers are bad*")

**Someone:** I don't agree that hypothesis "*most people buy food*" implies "*cash-transfers are b...*" (read more)

[[-](#)] **Eli Tyre** 6y 6 1

Or another way to put it, is that "someone" unskillfully said "the argument I offered is not a crux for me."

[[-](#)] **zulupineapple** 8y 7 0

Regarding solutions, I wish we could force people to explicitly answer questions. For example, suppose Someone's last reply said

**Someone:** Yes, really. But regardless, do you think that malnourished people don't deserve to buy a bit of food?

**Jacobian:** I don't actually think that. Now, why really?

And even better if we also forced people to ask specific questions with useful answers we could have

**Jacobian:** Why do you believe this? I thought most of the money was spent on quality-of-life things like metal roofs and better food.

**Someone:** I read it in... (read more)

[[-](#)] **Wei Dai** 8y 32 0

I generally liked this post, but have some caveats and questions.

Why does local validity work as well as it does in math? This post explains why it will lead you only to the truth, but why does it often lead to mathematical truths that we care about in a reasonable amount of time? In other words, why aren't most interesting math questions like  $P=NP$ , or how to win a game of chess?

Relying purely on local validity won't get you very far in playing chess, or life in general, and we instead have to frequently use intuitions. As Jan\_Kulveit [pointed out](#)<sup>o</sup>, it's generally not introspectively accessible why we have the intuitions that we do, so we sometimes make up bad arguments to explain them. So if someone points out the badness of an argument, it's important to adjust your credence downwards only as much as you considered the bad argument to be additional evidence on top of your intuitions.

Lastly, I can see the analogy between proofs and more general arguments, but I'm not sure about the third thing. Local validity. I mean I can see some surface similarities between them, but on

the third thing, law/coordination. I mean I can see some surface similarities between them, but are there also supposed to be deep correspondences between their underlying mechanisms? If so, I'm afraid I didn't get what they are.

[] **orthonormal** 8y 12 0

Relying purely on local validity won't get you very far in playing chess

The equivalent of local validity is just mechanically checking "okay, if I make this move, then they can make that move" for a bunch of cases. Which, first, is a major developmental milestone for kids learning chess. So we only think it "won't get you very far" because all the high-level human play explicitly or implicitly takes it for granted.

And secondly, it's pretty analogous to doing math; proving theorems is based on the ability to check the local validity of each step, but mathematicians aren't just brute-forcing their way to proofs. They have to develop higher-level heuristics, some of which are really hard to express in language, to suggest avenues, and then check local validity once they have a skeleton of some part of the argument. But if mathematicians stopped doing that annoying bit, well, then after a while you'll end up with another **crisis of analysis** when the brilliant intuitions are missing some tiny ingredient.

Local validity is an incredibly important part of any scientific discipline; the fact that it's not a part of most political discourse is merely a reflection that our society is at about the developmental level of a seven-year-old when it comes to political reasoning.

[] **Wei Dai** 8y 8 0

I suspect there may be a miscommunication here. To elaborate on "Relying purely on local validity won't get you very far in playing chess", what I had in mind is that if you decided to play a move only if you can prove that's it's the optimal move, you won't get very far, since we can't produce proofs of this form (even by using higher-level heuristics to guide us). Was your comment meant as a response to this point, or to a different interpretation of what I wrote?

[] **orthonormal** 8y 8 0

My comment was meant to explain what I understood Eliezer to be saying, because I think you had misinterpreted that. The OP is simply saying "don't give weight to arguments that are locally invalid, regardless of what else you like about them". Of course you need to use priors, heuristics, and intuitions in areas where you can't find an argument that carries you from beginning to end. But being able to think "oh, if I move there, then they can take my queen, and I don't see anything else good about that position, so let's not do that then" is a fair bit easier than proving your move optimal.

[] **Wei Dai** 8y 13 0

Oh, I see. I think your understanding of Eliezer makes sense, and the sentence you responded to wasn't really meant as an argument against the OP, but rather as setup for the point of that paragraph, which was "if someone points out the badness of an argument, it's important to adjust your credence downwards only as much as you considered the bad argument to be additional evidence on top of your intuitions."

To elaborate, I thought there was a possible mistake someone might make after reading the OP and wanted to warn against that. Specifically the mistake is that someone makes a bad

argument to explain an intuition, the badness of the argument is pointed out, they accept that and then give up their intuition or adjust their credence downward too much. This is not an issue for most people who haven't read the OP because they would just refuse to accept the badness of the argument.

(ETA: On second thought maybe I did initially misinterpret Eliezer and meant to argue against him, and am now forgetting that and giving a different motivation for what I wrote. In any case, I currently think your interpretation is correct, and what I wrote may still be valuable in that light. :)

[ - ] **zulupineapple** 8y 8 0

Why does local validity work as well as it does in math? <...> In other words, why aren't most interesting math questions like  $P=NP$ , or how to win a game of chess?

Why do you think that it works well? Are you sure most possible mathematical questions aren't exactly like  $P=NP$ , or worse? The set of "interesting" questions isn't representative of all questions, this set starts with " $2+2=?$ " and grows slowly, new questions become "interesting" only after old ones are answered. There is also some intuition about ... (read more)

[ - ] **Ben Pace** 6y 26 0 *Review for 2018 Review*

I think about this post a lot, and sometimes in conjunction with my own post on [common knowledge](#).

As well as it being a referent for when I think about fairness, it also ties in with how I think about LessWrong, Arbital and communal online endeavours for truth. The key line is:

For civilization to hold together, we need to make coordinated steps away from Nash equilibria in lockstep.

You can think of Wikipedia as being a set of communally editable web pages where the content of the page is constrained to be that which we can easily gain common knowledge of its truth. Wikipedia's information is only that which comes from [verifiable sources](#), which is how they solve this problem - all the editors don't have to get in a room and talk forever if there's a simple standard of truth. (I mean, they still do, but it would blow up to an impossible level if the standard were laxer than this.)

I understand a key part of the vision for [Arbital](#) was that, instead of the common standard being verifiable *facts*, it was instead to build a site around verifiable *steps of inference*, or alternatively phrased, *local validity*. This would allow us to walk through argument space together without knowing whether t  
... (read more)

[ - ] **bryjnar** 8y 21 0

The point that the Law needs to be *simple* and *local* so that humans can cope with it is also true of other domains. And this throws up an important constraint for people *designing* systems that humans are supposed to interact with: you must make it possible to reason simply and locally

...and we suppose to interact with, you make make it possible to reason simply, and reason, about them.

This comes up in programming (to a man with a nail everything looks like a hammer): good programming practice emphasises splitting programs up into small components that can be reasoned about in isolation. Modularity, compositionality, abstraction, etc. aside from their other benefits, make it possible to reason about code *locally*.

Of course, some people inexplicably believe that programs are mostly supposed to be consumed by computers, which have very different simplicity requirements and don't care much about locality. This can lead to programs that are very difficult for *humans* to consume.

Similarly, if you are writing a mathematical proof, it is good practice to try and split it up into small lemmas, transform the domain with definitions to make it simpler, and prove sub-components in isolation.

Interestingly, these days you can also write mathematical proofs to be consumed by a comput...  
(read more)

[[-](#)] **Zack\_M\_Davis** 6y 20 0 [Review for 2018 Review](#)

It strikes me as pedagogically unfortunate that sections i. and ii. (on arguments and proof-steps being locally valid) are part of the same essay as sections iii.–vi. (on what this has to do with the function of Law in Society). Had this been written in the Sequences-era, one would imagine this being (at least) two separate posts, and it would be nice to have a reference link for *just* the concept of argumentative local validity (which is *obviously* correct and important to have a name for, even if some of the speculations about Law in sections iii.–vi. turned out to be wrong).

[[-](#)] **Zvi** 6y 19 0 [Review for 2018 Review](#)

I find it deeply sad that many of us feel the need to frequently link to this article - I don't think I have ever done so, because if I need to explain local validity, then perhaps I'm talking to the wrong people? But certainly the ignoring of this principle has gotten more and more blatant and common over time since this post, so it's becoming less reasonable to assume that people understand such things. Which is super scary.

[[-](#)] **Raemon** 6y 3 0

I'm actually somewhat rolling to disbelieve on how frequently people link to the article (I guess because of my own experience which I'm now typical minding).

I haven't personally linked to this *because someone was failing at local validity*. What I've done is refer to this article while (attempting to) work towards a deeper, comprehensive culture that takes it seriously. What I found valuable about this was not "here's a thing that previously we didn't know and now we know it", it's "we were all sort of acting upon this belief, but not in a coordinated fashion. Now we have a good reference and common knowledge of it, which enables us to build off of it."

[[-](#)] **orthonormal** 6y 18 0 [Review for 2018 Review](#)

This post should be included in the Best-of-2018 compilation.

This is not only a good post, but one which cuts to the core of *what this community is about*. This site began not as a discussion of topics X, Y, and Z, but as a discussion of how to be... *less wrong* than the world around you (even/especially your own ingroup), and the difficulties this entails. Uncompromising honesty and self-skepticism are hard, and even though the best parts are a distillation of other parts of the Sequences, people need to be reminded more often than they need to be instructed.

[[-](#)] **Zack\_M\_Davis** 7y 16 0 *Nomination for 2018 Review*

I link this post (and use the phrase "local validity") a *lot* as a standard reference for the "This particular step of an argument either makes sense or doesn't make sense, independently of whether you agree with the conclusion of the broader argument" idea.

[[-](#)] **Jameson Quinn** 6y 12 0 *Review for 2018 Review*

I understand that this post seems wise to some people. To me, it seems like a series of tautologies on the surface, with an understructure of assumptions that are ultimately far more important and far more questionable. The basic assumption being made is that society-wide "memetic collapse" is a thing; the evidence given for this (even if you follow the links) is weak, and yet the attitude throughout is that further debate on this point is not worth our breath.

I am a co-author of statistics work with somebody whose standards of mathematical rigour are higher than mine. I often take intuitive leaps that she questions. There are three different common outcomes: one, that once we add the necessary rigour to my initial claim, it turns out to be right; two, that we recognize a mistake, but are able to fix things with some relatively minor claim to the original claim; and three, that I turn out to have been utterly wrong. So yes, it's good that she demands rigour, but I think that if she said "you are a Bad Person for making a locally-invalid argument" every time I made a leap, our collaboration would be less productive overall.

Overall, I don't object to the... (read more)

[[-](#)] **Hazard** 8y 12 0

The game-theoretic function of law can make following those simple rules feel like losing something, taking a step backward. You don't get to defect in the Prisoner's Dilemma, you don't get that delicious (5, 0) payoff instead of (3, 3). The law may punish one of your allies. You may be losing something according to your actual value function, which *feels like*° the law having an objectively bad immoral result. You may coherently hold that the universe is a worse place for an instance of the enforcement of a good law, relative to its counterfactual state if that law could be lifted in just that instance without affecting any other instances. Though this does require seeing that law as having a game-theoretic function as well as a moral function.

Growing up I've been firmly in the camp of "rules and the law are bullshit!". At the beginning of going down the rationality path, that meme was only bolstered by being able to see through people's inconsistent explanations for why things were. I've since moved away from that, but this post made me realize that my old "rules are bullshit" attitude is still operating in the background of my mi... (read more)

[[-](#)] **Vaniver** 7y\* 11 0 *Nomination for 2018 Review*

I think this post is worthy of inclusion basically because it contributes to solving this problem:

I've been musing recently about how a lot of the standard Code of the Light isn't really written down anywhere anyone can find.

That said, I think it's marred a little bit by reference to events and people in 2018 that might quickly pass from memory (as I think the main 3 names referred to have left public life, at least temporarily). This seems easy to fix with footnotes (which seems better than taking them out or otherwise making the point less specific).

[[-](#)] **Douglas\_Reay** 8y 7 0

- For civilization to hold together, we need to make coordinated steps away from Nash equilibria in lockstep. This requires general rules that are allowed to impose penalties on people we like or reward people we don't like. When people stop believing the general rules are being evaluated sufficiently fairly, they go back to the Nash equilibrium and civilization falls.

Two similar ideas:

There is a group evolutionary advantage for a society to support punishing those who defect from the social contract.

We get the worst democracy that we're willing... ([read more](#))

[[-](#)] **ChristianKI** 8y 7 0

I'm not sure what to make of this argument.

There's one ideal of disagreement where you focus on disagreeing with the central point and steelman what the other person says. We frequently advocate for that norm on LessWrong.

To me this post seems to point in the opposite direction. Instead of steelmanning a bad argument that's made, you are supposed to challenge the bad argument directly.

What kind of norm do we want to value on LessWrong?

[[-](#)] **Rob Bensinger** 8y 31 0

Eliezer wrote this in a private Facebook thread February 2017:

Reminder: Eliezer [and Holden](#) are both on record as saying that "steelmanning" people is bad and you should stop doing it.

As Holden says, if you're trying to understand someone or you have any credence at all that they have a good argument, focus on passing their Ideological Turing Test.

"Steelmanning" usually ends up as weakmanning by comparison. If they don't in fact have a good argument, it's falsehood to pretend they do. If you want to try to make a

genuine effort to think up better arguments yourself because they might exist, don't drag the other person into it.

And he **FB-commented** on Ozy's **Against Steelmanning** in August 2016:

Be it clear: Steelmanning is not a tool of understanding and communication. The communication tool is the Ideological Turing Test. "Steelmanning" is what you do to avoid the equivalent of dismissing AGI after reading a media argument. It usually indicates that you think you're talking to somebody as hapless as the media.

The exception to this rule is when you communicate, "Well, on my assumptions, the plausible thing that sounds most

... (read more)

[] **tristanm** 8y 3 0

It seems like in the vast majority of conversations, we find ourselves closer to the "exposed to the Deepak Chopra version of quantum mechanics and haven't seen the actual version yet" situation than we do to the "Arguing with someone who is far less experienced and knowledgeable than you are on this subject." In the latter case, it's easy to see why steelmanning would be counterproductive. If you're a professor trying to communicate a difficult subject to a student, and the student is having trouble understanding your position, it's u... (read more)

[] **tristanm** 8y 1 0

I don't see him as arguing against steelmanning. But the opposite of steelmanning isn't arguing against an idea directly. You've got to be able to steelman an opponent's argument well in order to argue against it well too, or perhaps determine that you agree with it. In any case, I'm not sure how to read a case for locally valid argumentation steps as being in favor of not doing this. Wouldn't it help you understand how people arrive at their conclusions?

[] **ChristianKI** 8y 3 0

There are plenty of times where someone writes a LessWrong post and while I do agree with the central point of the post I disagree with a noncentral part of the post.

A person might use some historical example and I disagree with the example. In those cases it's for me an open question whether or not it's useful to write the comment that disagrees or whether that's bad for LW. It might be bad because people feel like they are getting noncentral feedback and that discourages them.

[] **zulupineapple** 8y 6 0

From the linked fb post on memetic collapse:

We're looking at a collapse of reference to expertise

I think this much might be true, and I think there is a single meme responsible for a large part of the collapse. That is the "*think for yourself. don't be a sheeb*" meme. Problem is that everyone likes

this meme. I'm sure it's popular on LW too. And, of course, there are healthy ways to "*think for yourself*". But if you tell someone not to be a sheep, and they start rejecting reasonable experts in favor of their intuitions... (read more)

[[-](#)] **Arthur Milchior** 5y 4 0

A small note, which would probably have been better before it get published. For someone not following US politics, the story of Doug Jones is hard to follow as I don't have any context about it. My first reading suggest that some people would have wanted to expel a senator, leaving the senate with one less member on the democratic side. But it does not seems to make sens, unless some party have the power to expell a senator from the whole senate

[[-](#)] **Ruby** 7y 4 0 *Nomination for 2018 Review*

This seems like such a core idea. I can't recall employing it, but it's stuck in my mind as distinction that makes our community especially valuable and helps me understand that the lack of this is what's behind so much insanity I see in others. So I guess I don't feel like I "use", but it has sunk deep into my models of how people reason.

[[-](#)] **ChristianKI** 8y 4 0

Eric Ries of the Lean Startup frequently says that every company should have a chief officer of innovation because unless it's someone's job to focus on innovation that function isn't done well.

In a similar way, every government should have a minister of deregulation that oversees a bunch of professionals that work on proposing ways to make the government simpler.

Deregulation and we simply need to make it more of a political or policy priority to invest enough manpower into working on it.

[[-](#)] **Eli Tyre** 6y 3 0

The notion that you can "be fair to one side but not the other", that what's called "fairness" is a kind of favor you do for people you like, says that even the *instinctive* sense people had of law-as-game-theory is being lost in the modern memetic collapse. People are being exposed to so many social-media-viral depictions of the Other Side defecting, and viewpoints exclusively from Our Side without any leavening of any other viewpoint that might ask for a game-theoretic compromise, that they're losing the ability to app

... (read more)

[[-](#)] **Adam Zerner** 3y 2 0

I'd much rather buy a used car from the second person than the first person. I think I'd pay at least a 5% price premium.

I really like this. Illustrating with "who I'd rather buy a used car from". Cool.

[–] **DPiegrass** 5y 2 0

Lord Kelvin's careful and multiply-supported lines of reasoning arguing that the Earth could not possibly be so much as a hundred million years old, all failed simultaneously in a surprising way because that era didn't know about nuclear reactions.

I'm told that the biggest reason **Kelvin was wrong** was that, for many years, no one thought about there being a molten interior subject to convection:

Perry's [1895?] calculation shows that if the Earth has a conducting lid of 50 kilometers' thickness, with a perfectly convecting fluid underneath, then the me

... (read more)

[–] **gustaf** 18d 1 0

If, like me, you want to link someone to just the concept of Local Validity, then consider **Validity (LessWrong Wiki)**°

[–] **Ysamuels** 🌱 3mo\* 1 0

I think you're thinking of a combination of the Biblical stories of Samuel 1, 14 with Jonathan and Saul, and Judges 11 with Jeptath's daughter (who'd have thought my religious education would ever come in handy on lesswrong!)

[–] **mike\_hawke** 3y 1 0

Is this different than logical validity? If not, how do they relate?

[https://en.wikipedia.org/wiki/Validity\\_\(logic\)](https://en.wikipedia.org/wiki/Validity_(logic))

<https://www.lesswrong.com/tag/valid-argument>°

I could believe that adding the word "local" might help people communicate the concept more clearly, but I'm curious if it's doing anything else here.

[–] **Ben Pace** 8y 1 0

I've curated this post for these reasons:

- A key social concept was cached out in a mathematical description, which is really useful. Connecting practical knowledge to our models of reality that are most reliable (mathematics) is a central rationalist skill.
- The post (and comments) helped point me to the *feeling* of fairness (which feels close to me to the feeling of verifying that a technical explanation has been offered), which is a valuable thing to have a concept-handle ('local validity') for.
- As with the rest of Eliezer's posts, it was

... (read more)

[ - ] **TheWakalix** 8y 1 0

There will be scorching days even if the climate is slowly cooling. It's really not a significant amount of evidence - certainly not worth talking about. It's not worth the air you would talk it into.

[ - ] **Said Achmiz** 8y 1 0

There are weaker and stronger versions of this attribute. Some people will think to themselves, "Well, it's important to use only valid arguments... but there was a sustained pattern of record highs worldwide over multiple years which does count as evidence, and that particular very hot day was a part of that pattern, so it's valid evidence for global warming." Other people will think to themselves, "I'd roll my eyes at someone who offers a single very cold day as an argument that global warming is false. So it can't be okay to use a single very hot day to

... (read more)

[ - ] **Eliezer Yudkowsky** 8y 16 0

There will be a single very cold day occasionally regardless of whether global warming is true or false. Anyone who knows the phrase "modus tollens" ought to know that. That said, if two unenlightened ones are arguing back and forth in all sincerity by telling each other about the hot versus cold days they remember, neither is being dishonest, but both are making invalid arguments. But this is not the scenario offered in the original, which concerns somebody who does possess the mental resources to know better, but is tempted to rationalize in order to reach the more agreeable conclusion. They feel a little pressure in their head when it comes to deciding which argument to accept. If a judge behaved thusly in sentencing a friend or an enemy, would we not consider them morally deficient in their duty as a judge? There is a level of unconscious ignorance that renders an innocent entirely blameless; somebody who possesses the inner resources to have the first intimation that one hot day is a bad argument for global warming is past that level.

[ - ] **Said Achmiz** 8y 3 0

That said, if two unenlightened ones are arguing back and forth in all sincerity by telling each other about the hot versus cold days they remember, neither is being dishonest, but both are making invalid arguments.

Yes, that is what I was saying.

But this is not the scenario offered in the original ...

The (apparent) reasoning I am challenging, Eliezer, is this: "Alice is making an invalid argument for side A" -> "therefore Alice would *not* make an invalid argument for side B". This seems faulty. You seem to be taking someone's use of invalid reasonin... (read more)

[ - ] **rk** 8y 7 0

The (apparent) reasoning I am challenging, Eliezer, is this: "Alice is making an invalid

The (apparent) reasoning I am challenging, Eliezer, is this. Alice is making an invalid argument for side A" -> "therefore Alice would *not* make an invalid argument for side B". This seems faulty.

As some anecdotal data, I know several people who have done this precisely for global warming. I suspect Eliezer has spoken to them as well, because the standard response when you point out that they're using the reverse of the "day of cold weather" argument they've derided is:

Well, it's important to use only valid arguments... but there was a sustaine  
... (read more)

[Moderation Log](#)

## More from Eliezer Yudkowsky

|     |                                                        |                   |     |     |
|-----|--------------------------------------------------------|-------------------|-----|-----|
| 355 | The Owned Ones                                         | Eliezer Yudkowsky | 21d | 50  |
| 203 | A relatively brief explanation of Boltzmann Brains     | Eliezer Yudkowsky | 17d | 149 |
| 365 | Irretrievability; or, Murphy's Curse of Oneshotness... | Eliezer Yudkowsky | 18d | 132 |

[View more](#)

## Curated and popular this week

|     |                                                            |               |     |     |
|-----|------------------------------------------------------------|---------------|-----|-----|
| 153 | Empowerment, corrigibility, etc. are simple abstractio...  | Steven Byrnes | 1d  | 62  |
| 255 | Bad Problems Don't Stop Being Bad Because Somebody's Wr... | Linch         | 6d  | 73  |
| 322 | Women should be able to open things                        | KatjaGrace    | 12d | 130 |

