Best of LessWrong: August 2012

1. The noncentral fallacy - the worst argument in the world?
2. Bayes for Schizophrenics: Reasoning in Delusional Disorders
3. What are the optimal biases to overcome?
4. AI timeline predictions: are we getting better?
5. "Epiphany addiction"
6. Why Don't People Help Others More?
7. Solving the two envelopes problem
8. Who Wants To Start An Important Startup?
9. Self-skepticism: the first principle of rationality
10. A model of UDT with a concrete prior over logical statements
11. The High Impact Network (THINK) - Launching Now
12. Decision Theories, Part 3.5: Halt, Melt and Catch Fire
13. An angle of attack on Open Problem #1
14. Natural Laws Are Descriptions, not Rules

Best of LessWrong: August 2012

1. The noncentral fallacy - the worst argument in the world?
2. Bayes for Schizophrenics: Reasoning in Delusional Disorders
3. What are the optimal biases to overcome?
4. AI timeline predictions: are we getting better?
5. "Epiphany addiction"
6. Why Don't People Help Others More?
7. Solving the two envelopes problem
8. Who Wants To Start An Important Startup?
9. Self-skepticism: the first principle of rationality
10. A model of UDT with a concrete prior over logical statements
11. The High Impact Network (THINK) - Launching Now
12. Decision Theories, Part 3.5: Halt, Melt and Catch Fire
13. An angle of attack on Open Problem #1
14. Natural Laws Are Descriptions, not Rules

The noncentral fallacy - the worst argument in the world?

Related to: Leaky Generalizations, Replace the Symbol With The Substance, Sneaking In Connotations

David Stove once ran a contest to find the Worst Argument In The World, but he awarded the prize to his own entry, and one that shored up his politics to boot. It hardly seems like an objective process.

If he can unilaterally declare a Worst Argument, then so can I. I declare the Worst Argument In The World to be this: "X is in a category whose archetypal member gives us a certain emotional reaction. Therefore, we should apply that emotional reaction to X, even though it is not a central category member."

Call it the Noncentral Fallacy. It sounds dumb when you put it like that. Who even does that, anyway?

It sounds dumb only because we are talking soberly of categories and features. As soon as the argument gets framed in terms of *words*, it becomes so powerful that somewhere between many and most of the bad arguments in politics, philosophy and culture take some form of the noncentral fallacy. Before we get to those, let's look at a simpler example.

Suppose someone wants to build a statue honoring Martin Luther King Jr. for his nonviolent resistance to racism. An opponent of the statue objects: "But Martin Luther King was a *criminal*!"

Any historian can confirm this is correct. A criminal is technically someone who breaks the law, and King knowingly broke a law against peaceful anti-segregation protest - hence his famous Letter from Birmingham Jail.

But in this case calling Martin Luther King a criminal is the noncentral. The archetypal criminal is a mugger or bank robber. He is driven only by greed, preys on the innocent, and weakens the fabric of society. Since we don't like these things, calling someone a "criminal" naturally lowers our opinion of them.

The opponent is saying "Because you don't like criminals, and Martin Luther King is a criminal, you should stop liking Martin Luther King." But King doesn't share the important criminal features of being driven by greed, preying on the innocent, or weakening the fabric of society that made us dislike criminals in the first place. Therefore, even though he is a criminal, there is no reason to dislike King.

This all seems so nice and logical when it's presented in this format. Unfortunately, it's also one hundred percent contrary to instinct: the urge is to respond "Martin Luther King? A criminal? No he wasn't! You take that back!" This is why the noncentral is so successful. As soon as you do that you've fallen into their trap. Your argument is no longer about whether you should build a statue, it's about whether King was a criminal. Since he was, you have now lost the argument.

Ideally, you should just be able to say "Well, King was the good kind of criminal." But

that seems pretty tough as a debating maneuver, and it may be even harder in some of the cases where the noncentral Fallacy is commonly used.

Now I want to list some of these cases. Many will be political[1], for which I apologize, but it's hard to separate out a bad argument from its specific instantiations. None of these examples are meant to imply that the position they support is wrong (and in fact I myself hold some of them). They only show that certain particular arguments for the position are flawed, such as:

"**Abortion is *murder*!"** The archetypal murder is Charles Manson breaking into your house and shooting you. This sort of murder is bad for a number of reasons: you prefer not to die, you have various thoughts and hopes and dreams that would be snuffed out, your family and friends would be heartbroken, and the rest of society has to live in fear until Manson gets caught. If you define murder as "killing another human being", then abortion is technically murder. But it has none of the downsides of murder Charles Manson style. Although you can criticize abortion for many reasons, insofar as "abortion is murder" is an invitation to apply one's feelings in the Manson case directly to the abortion case, it ignores the latter's lack of the features that generated those intuitions in the first place[2].

"**Genetic engineering to cure diseases is *eugenics*!**" Okay, you've got me there: since eugenics means "trying to improve the gene pool" that's clearly right. But what's wrong with eugenics? "What's wrong with eugenics? Hitler did eugenics! Those unethical scientists in the 1950s who sterilized black women without their consent did eugenics!" "And what was wrong with what Hitler and those unethical scientists did?" "What do you mean, what was wrong with them? Hitler killed millions of people! Those unethical scientists ruined people's lives." "And does using genetic engineering to cure diseases kill millions of people, or ruin anyone's life?" "Well...not really." "Then what's wrong with it?" "It's *eugenics*!"

"Evolutionary psychology is *sexist*!" If you define "sexist" as "believing in some kind of difference between the sexes", this is true of at least some evo psych. For example, Bateman's Principle states that in species where females invest more energy in producing offspring, mating behavior will involve males pursuing females; this posits a natural psychological difference between the sexes. "Right, so you admit it's sexist!" "And why exactly is sexism bad?" "Because sexism claims that men are better than women and that women should have fewer rights!" "Does Bateman's principle claim that men are better than women, or that women should have fewer rights?" "Well...not really." "Then what's wrong with it?" "It's *sexist!*"

A second, subtler use of the noncentral fallacy goes like this: "X is in a category whose archetypal member gives us an emotional reaction. Therefore, we should apply that same emotional reaction to X even if X gives some benefit that outweighs the harm."

"Capital punishment is *murder*!" Charles Manson-style murder is solely harmful. This kind of murder produces really strong negative feelings. The proponents of capital punishment believe that it might decrease crime, or have some other attending benefits. In other words, they believe it's "the good kind of murder"[3], just like the introductory example concluded that Martin Luther King was "the good kind of criminal". But since normal murder is so taboo, it's really hard to take the phrase "the good kind of murder" seriously, and just mentioning the word "murder" can call up exactly the same amount of negative feelings we get from the textbook example.

"**Affirmative action is *racist*!**" True if you define racism as "favoring certain people based on their race", but once again, our immediate negative reaction to the archetypal example of racism (the Ku Klux Klan) cannot be generalized to an immediate negative reaction to affirmative action. Before we generalize it, we have to check first that the problems that make us hate the Ku Klux Klan (violence, humiliation, divisiveness, lack of a meritocratic society) are still there. Then, even if we do find that some of the problems persist (like disruption of meritocracy, for example) we have to prove that it doesn't produce benefits that outweigh these harms.

"**Taxation is *theft*!**" True if you define theft as "taking someone else's money regardless of their consent", but though the archetypal case of theft (breaking into someone's house and stealing their jewels) has nothing to recommend it, taxation (arguably) does. In the archetypal case, theft is both unjust and socially detrimental. Taxation keeps the first disadvantage, but arguably subverts the second disadvantage if you believe being able to fund a government has greater social value than leaving money in the hands of those who earned it. The question then hinges on the relative importance of these disadvantages. Therefore, you can't dismiss taxation without a second thought just because you have a natural disgust reaction to theft in general. You would also have to prove that the supposed benefits of this form of theft don't outweigh the costs.

Now, because most arguments are rapid-fire debate-club style, sometimes it's still useful to say "Taxation isn't theft!" At least it beats saying "Taxation is theft but nevertheless good", then having the other side say "Apparently my worthy opponent thinks that theft can be good; we here on this side would like to bravely take a stance *against* theft", and then having the moderator call time before you can explain yourself. If you're in a debate club, do what you have to do. But if you have the luxury of philosophical clarity, you would do better to forswear the Dark Arts and look a little deeper into what's going on.

Are there ever cases in which this argument pattern can be useful? Yes. For example, it may be a groping attempt to suggest a Schelling fence; for example, a principle that one must never commit theft even when it would be beneficial because that would make it harder to distinguish and oppose the really bad kinds of theft. Or it can be an attempt to spark conversation by pointing out a potential contradiction: for example "Have you noticed that taxation really does contain some of the features you dislike about more typical instances of theft? Maybe you never even thought about that before? Why do your moral intuitions differ in these two cases? Aren't you being kind of hypocritical?" But this usage seems pretty limited - once your interlocutor says "Yes, I considered that, but the two situations are different for reasons X, Y, and Z" the conversation needs to move on; there's not much point in continuing to insist "But it's *theft*!"

But in most cases, I think this is more of an *emotional* argument, or even an argument from "You would look silly saying that". You really *can't* say "Oh, he's the good kind of criminal", and so if you have a potentially judgmental audience and not much time to explain yourself, you're pretty trapped. You have been forced to round to the archetypal example of that word and subtract exactly the information that's most relevant.

But in all other cases, the proper response to being asked to subtract relevant information is "No, why should I?" - and that's why this is the worst argument in the world.

Footnotes

1: On advice from the community, I have deliberately included three mostly-liberal examples and three-mostly conservative examples, so save yourself the trouble of counting them up and trying to speculate on this article's biases.

2: This should be distinguished from deontology, the belief that there is some provable moral principle about how you can never murder. I don't think this is *too* important a point to make, because only a tiny fraction of the people who debate these issues have thought that far ahead, and also because my personal and admittedly controversial opinion is that much of deontology is just an attempt to formalize and justify this fallacy.

3: Some people "solve" this problem by saying that "murder" only refers to "non-lawful killing", which is exactly as creative a solution as redefining "criminal" to mean "person who breaks the law and is not Martin Luther King." Identifying the noncentral fallacy is a more complete solution: for example, it covers the related (mostly sarcastic) objection that "imprisonment is kidnapping".

4: EDIT 8/2013: I've edited this article a bit after getting some feedback and complaints. In particular I tried to remove some LW jargon which turned off some people who were being linked to this article but were unfamiliar with the rest of the site.

5: EDIT 8/2013: The other complaint I kept getting is that this is an uninteresting restatement of some other fallacy (no one can agree which, but poisoning the well comes up particularly often). The question doesn't seem too interesting to me - I never claimed particular originality, a lot of fallacies blend into each other, and the which-fallacy-is-which game isn't too exciting anyway - but for the record I don't think it is. Poisoning the well is a presentation of two different facts, such as "Martin Luther King was a plagiarist...oh, by the way, what do you think of Martin Luther King's civil rights policies?" It may have no relationship to categories, and it's usually something someone else does to you as a conscious rhetorical trick. Noncentral fallacy is presenting a single fact, but using category information to frame it in a misleading way - and it's often something people do to themselves. The above plagiarism example of poisoning the well is *not* noncentral fallacy. If you think this essay is about bog-standard poisoning the well, then either there is an alternative meaning to poisoning the well I'm not familiar with, or you are missing the point.

Bayes for Schizophrenics: Reasoning in Delusional Disorders

Related to: The Apologist and the Revolutionary, Dreams with Damaged Priors

Several years ago, I posted about V.S. Ramachandran's 1996 theory explaining anosognosia through an "apologist" and a "revolutionary".

Anosognosia, a condition in which extremely sick patients mysteriously deny their sickness, occurs during right-sided brain injury but not left-sided brain injury. It can be extraordinarily strange: for example, in one case, a woman whose left arm was paralyzed insisted she could move her left arm just fine, and when her doctor pointed out her immobile arm, she claimed that was her daughter's arm even though it was obviously attached to her own shoulder. Anosognosia can be temporarily alleviated by squirting cold water into the patient's left ear canal, after which the patient suddenly realizes her condition but later loses awareness again and reverts back to the bizarre excuses and confabulations.

Ramachandran suggested that the left brain is an "apologist", trying to justify existing theories, and the right brain is a "revolutionary" which changes existing theories when conditions warrant. If the right brain is damaged, patients are unable to change their beliefs; so when a patient's arm works fine until a right-brain stroke, the patient cannot discard the hypothesis that their arm is functional, and can only use the left brain to try to fit the facts to their belief.

In the almost twenty years since Ramachandran's theory was published, new research has kept some of the general outline while changing many of the specifics in the hopes of explaining a wider range of delusions in neurological and psychiatric patients. The newer model acknowledges the left-brain/right-brain divide, but adds some new twists based on the Mind Projection Fallacy and the brain as a Bayesian reasoner.

INTRODUCTION TO DELUSIONS

Strange as anosognosia is, it's only one of several types of delusions, which are broadly categorized into polythematic and monothematic. Patients with polythematic delusions have multiple unconnected odd ideas: for example, the famous schizophrenic game theorist John Nash believed that he was defending the Earth from alien attack, that he was the Emperor of Antarctica, *and* that he was the left foot of God. A patient with a monothematic delusion, on the other hand, usually only has one odd idea. Monothematic delusions vary less than polythematic ones: there are a few that are relatively common across multiple patients. For example:

In the Capgras delusion, the patient, usually a victim of brain injury but sometimes a schizophrenic, believes that one or more people close to her has been replaced by an identical imposter. For example, one male patient expressed the worry that his wife was actually someone else, who had somehow contrived to exactly copy his wife's appearance and mannerisms. This delusion sounds harmlessly hilarious, but it can get very ugly: in at least one case, a patient got so upset with the deceit that he murdered the hypothesized imposter - actually his wife.

The Fregoli delusion is the opposite: here the patient thinks that random strangers she meets are actually her friends and family members in disguise. Sometimes everyone may be the same person, who must be as masterful at quickly changing costumes as the famous Italian actor Fregoli (inspiring the condition's name).

In the Cotard delusion, the patient believes she is dead. Cotard patients will neglect personal hygiene, social relationships, and planning for the future - as the dead have no need to worry about such things. Occasionally they will be able to describe in detail the "decomposition" they believe they are undergoing.

Patients with all these types of delusions[1] - as well as anosognosiacs - share a common feature: they usually have damage to the right frontal lobe of the brain (including in schizophrenia, where the brain damage is of unknown origin and usually generalized, but where it is still possible to analyze which areas are the most abnormal). It would be nice if a theory of anosognosia also offered us a place to start explaining these other conditions, but this Ramachandran's idea fails to do. He posits a problem with belief shift: going from the originally correct but now obsolete "my arm is healthy" to the updated "my arm is paralyzed". But these other delusions cannot be explained by simple failure to update: delusions like "the person who appears to be my wife is an identical imposter" *never* made sense. We will have to look harder.

ABNORMAL PERCEPTION: THE FIRST FACTOR

Coltheart, Langdon, and McKay posit what they call the "two-factor theory" of delusion. In the two-factor theory, one problem causes an abnormal perception, and a second problem causes the brain to come up with a bizarre instead of a reasonable explanation.

Abnormal perception has been best studied in the Capgras delusion. A series of experiments, including some by Ramachandran himself, demonstrate that Capgras patients lack a skin conductance response (usually used as a proxy of emotional reaction) to familiar faces. This meshes nicely with the brain damage pattern in Capgras, which seems to involve the connection between the face recognition areas in the temporal lobe and the emotional areas in the limibic system. So although the patient can recognize faces, and can feel emotions, the patient cannot feel emotions related to recognizing faces.

The older "one-factor" theories of delusion stopped here. The patient, they said, knows that his wife looks like his wife, but he doesn't feel any emotional reaction to her. If it was really his wife, he would feel something - love, irritation, whatever - but he feels only the same blankness that would accompany seeing a stranger. Therefore (the one-factor theory says) his brain gropes for an explanation and decides that she really is a stranger. Why does this stranger look like his wife? Well, she must be wearing a very good disguise.

One-factor theories also do a pretty good job of explaining many of the remaining monothematic delusions. A 1998 experiment shows that Cotard delusion sufferers have a globally decreased autonomic response: that is, nothing really makes them feel much of anything - a state consistent with being dead. And anosognosiacs have lost not only the nerve connections that would allow them to move their limbs, but the nerve connections that would send distress signals and even the connections that would send back "error messages" if the limb failed to move correctly - so the brain

gets data that everything is fine.

The basic principle behind the first factor is "Assume that reality is such that my mental states are justified", a sort of Super Mind Projection Fallacy.

Although I have yet to find an official paper that says so, I think this same principle also explains many of the more typical schizophrenic delusions, of which two of the most common are delusions of grandeur and delusions of persecution. Delusions of grandeur are the belief that one is extremely important. In pop culture, they are typified by the psychiatric patient who believes he is Jesus or Napoleon - I've never met any Napoleons, but I know several Jesuses and recently worked with a man who thought he was Jesus and John Lennon at the same time. Here the first factor is probably an elevated mood (working through a miscalibrated sociometer). "Wow, I feel like I'm really awesome. In what case would I be justified in thinking so highly of myself? Only if I were Jesus and John Lennon at the same time!" A similar mechanism explains delusions of persecution, the classic "the CIA is after me" form of disease. We apply the Super Mind Projection Fallacy to a garden-variety anxiety disorder: "In what case would I be justified in feeling this anxious? Only if people were constantly watching me and plotting to kill me. Who could do that? The CIA."

But despite the explanatory power of the Super Mind Projection Fallacy, the one-factor model isn't enough.

ABNORMAL BELIEF EVALUATION: THE SECOND FACTOR

The one-factor model requires people to be really stupid. Many Capgras patients were normal intelligent people before their injuries. Surely they wouldn't leap straight from "I don't feel affection when I see my wife's face" to "And therefore this is a stranger who has managed to look exactly like my wife, sounds exactly like my wife, owns my wife's clothes and wedding ring and so on, and knows enough of my wife's secrets to answer any question I put to her exactly like my wife would." The lack of affection vaguely supports the stranger hypothesis, but the prior for the stranger hypothesis is so low that it should never even enter consideration (remember this phrasing: it will become important later.) Likewise, we've all felt really awesome at one point or another, but it's never occurred to most of us that maybe we are simultaneously Jesus and John Lennon.

Further, most psychiatric patients with the deficits involved don't develop delusions. People with damage to the ventromedial area suffer the same disconnection between face recognition and emotional processing as Capgras patients, but they don't draw any unreasonable conclusions from it. Most people who get paralyzed don't come down with anosognosia, and most people with mania or anxiety don't think they're Jesus or persecuted by the CIA. What's the difference between these people and the delusional patients?

The difference is the right dorsolateral prefrontal cortex, an area of the brain strongly associated with delusions. If whatever brain damage broke your emotional reactions to faces or paralyzed you or whatever spared the RDPC, you are unlikely to develop delusions. If your brain damage also damaged this area, you are correspondingly more likely to come up with a weird explanation.

In his first papers on the subject, Coltheart vaguely refers to the RDPC as a "belief evaluation" center. Later, he gets more specific and talks about its role in Bayesian updating. In his chronology, a person damages the connection between face

recognition and emotion, and "rationally" concludes the Capgras hypothesis. In his model, even if there's only a 1% prior of your spouse being an imposter, if there's a 1000 times greater likelihood of you not feeling anything toward an imposter than to your real spouse, you can "rationally" come to believe in the delusion. In normal people, this rational belief then gets worn away by updating based on evidence: the imposter seems to know your spouse's personal details, her secrets, her email passwords. In most patients, this is sufficient to have them update back to the idea that it is really their spouse. In Capgras patients, the damage to the RDPC prevents updating on "exogenous evidence" (for some reason, the endogenous evidence of the lack of emotion itself still gets through) and so they maintain their delusion.

This theory has some trouble explaining why patients are still able to update about other situations, but Coltheart speculates that maybe the belief evaluation system is weakened but not totally broken, and can deal with anything except the ceaseless stream of contradictory endogenous information.

EXPLANATORY ADEQUACY BIAS

McKay makes an excellent critique of several questionable assumptions of this theory.

First, is the Capgras hypothesis *ever* plausible? Coltheart et al pretend that the prior is 1/100, but this implies that there is a base rate of your spouse being an imposter one out of every hundred times you see her (or perhaps one out of every hundred people has a fake spouse) either of which is preposterous. No reasonable person could entertain the Capgras hypothesis even for a second, let alone for long enough that it becomes their working hypothesis and develops immunity to further updating from the broken RDPC.

Second, there's no evidence that the ventromedial patients - the ones who lose face-related emotions but don't develop the Capgras delusion - once had the Capgras delusion but then successfully updated their way out of it. They just never develop the delusion to begin with.

McKay keeps the Bayesian model, but for him the second factor is not a deficit in updating in general, but a deficit in the use of priors. He lists two important criteria for reasonable belief: "explanatory adequacy" (what standard Bayesians call the likelihood ratio; the new data must be more likely if the new belief is true than if it is false) and "doxastic conservativism" (what standard Bayesians call the prior; the new belief must be reasonably likely to begin with given everything else the patient knows about the world).

Delusional patients with damage to their RDPC lose their ability to work with priors and so abandon all doxastic conservativism, essentially falling into a what we might term the Super Base Rate Fallacy. For them the only important criterion for a belief is explanatory adequacy. So when they notice their spouse's face no longer elicits any emotion, they decide that their spouse is not really their spouse at all. This does a great job of explaining the observed data - maybe the best job it's possible for an explanation to do. Its only minor problem is that it has a stupendously low prior, and this doesn't matter because they are no longer able to take priors into account.

This also explains why the delusional belief is impervious to new evidence. Suppose the patient's spouse tells personal details of their honeymoon that no one else could possibly know. There are several possible explanations: the patient's spouse really is the patient's spouse, or (says the left-brain Apologist) the patient's spouse is an alien

who was able to telepathically extract the relevant details from the patient's mind. The telepathic alien imposter hypothesis has great explanatory adequacy: it explains why the person looks like the spouse (the alien is a very good imposter), why the spouse produces no emotional response (it's not the spouse at all) and why the spouse knows the details of the honeymoon (the alien is telepathic). The "it's really your spouse" explanation only explains the first and the third observations. Of course, we as sane people know that the telepathic alien hypothesis has a very low base rate plausibility because of its high complexity and violation of Occam's Razor, but these are exactly the factors that the RDPC-damaged[2] patient can't take into account. Therefore, the seemingly convincing new evidence of the spouse's apparent memories only suffices to help the delusional patient infer that the imposter is telepathic.

The Super Base Rate Fallacy can explain the other delusional states as well. I recently met a patient who was, indeed, convinced the CIA were after her; of note she also had extreme anxiety to the point where her arms were constantly shaking and she was hiding under the covers of her bed. CIA pursuit is probably the best possible reason to be anxious; the only reason we don't use it more often is how few people are really pursued by the CIA (well, as far as we know). My mentor warned me not to try to argue with the patient or convince her that the CIA wasn't really after her, as (she said from long experience) it would just make her think I was in on the conspiracy. This makes sense. "The CIA is after you and your doctor is in on it" explains both anxiety and the doctor's denial of the CIA very well; "The CIA is not after you" explains only the doctor's denial of the CIA. For anyone with a pathological inability to handle Occam's Razor, the best solution to a challenge to your hypothesis is always to make your hypothesis more elaborate.

OPEN QUESTIONS

Although I think McKay's model is a serious improvement over its predecessors, there are a few loose ends that continue to bother me.

"You have brain damage" is also a theory with perfect explanatory adequacy. If one were to explain the Capgras delusion to Capgras patients, it would provide just as good an explanation for their odd reactions as the imposter hypothesis. Although the patient might not be able to appreciate its decreased complexity, they should at least remain indifferent between the two hypotheses. I've never read of any formal study of this, but given that someone must have tried explaining the Capgras delusion to Capgras patients I'm going to assume it doesn't work. Why not?

Likewise, how come delusions are so specific? It's impossible to convince someone who thinks he is Napoleon that he's really just a random non-famous mental patient, but it's also impossible to convince him he's Alexander the Great (at least I think so; I don't know if it's ever been tried). But him being Alexander the Great is also consistent with his observed data and his deranged inference abilities. Why decide it's the CIA who's after you, and not the KGB or Bavarian Illuminati?

Why is the failure so often limited to failed inference from mental states? That is, if a Capgras patient sees it is raining outside, the same process of base rate avoidance that made her fall for the Capgras delusion ought to make her think she's been transported to ther rainforest or something. This happens in polythematic delusion patients, where anything at all can generate a new delusion, but not those with monothematic delusions like Capgras. There must be some fundamental difference between how one draws inferences from mental states versus everything else.

This work also raises the question of whether one can one consciously use System II Bayesian reasoning to argue oneself out of a delusion. It seems improbable, but I recently heard about an n=1 personal experiment of a rationalist with schizophrenia who used successfully used Bayes to convince themselves that a delusion (or possibly hallucination; the story was unclear) was false. I don't have their permission to post their story here, but I hope they'll appear in the comments.

FOOTNOTES

1: I left out discussion of the Alien Hand Syndrome, even though it was in my sources, because I believe it's more complicated than a simple delusion. There's some evidence that the alien hand actually does move independently; for example it will sometimes attempt to thwart tasks that the patient performs voluntarily with their good hand. Some sort of "split brain" issues seem like a better explanation than simple Mind Projection.

2: The right dorsolateral prefrontal cortex also shows up in dream research, where it tends to be one of the parts of the brain shut down during dreaming. This provides a reasonable explanation of why we don't notice our dreams' implausibility while we're dreaming them - and Eliezer specifically mentions he can't use priors correctly in his dreams. It also highlights some interesting parallels between dreams and the monothematic delusions. For example, the typical "And then I saw my mother, but she was also somehow my fourth grade teacher at the same time" effect seems sort of like Capgras and Fregoli. Even more interestingly, the RDPC gets switched on during lucid dreaming, providing an explanation of why lucid dreamers are able to reason normally in dreams. Because lucid dreaming also involves a sudden "switching on" of "awareness", this makes the RDPC a good target area for consciousness research.

What are the optimal biases to overcome?

If you're interested in learning rationality, where should you start? Remember, instrumental rationality is about making decisions that get you what you want -- surely there are some lessons that will help you more than others.

You might start with the most famous ones, which tend to be the ones popularized by Kahneman and Tversky. But K&T were academics. They weren't trying to help people be more rational, they were trying to prove to other academics that people *were* irrational. The result is that they focused not on the most important biases, but the ones that were easiest to prove.

Take their famous anchoring experiment, in which they showed the spin of a roulette wheel affected people's estimates about African countries. The idea wasn't that roulette wheels causing biased estimates was a huge social problem; it was that no academic could possibly argue that this behavior was somehow rational. They thereby scored a decisive blow for psychology against economists claiming we're just rational maximizers.

Most academic work on irrationality has followed in K&T's footsteps. And, in turn, much of the stuff done by LW and CFAR has followed in the footsteps of this academic work. So it's not hard to believe that LW types are good at avoiding these biases and thus do well on the psychology tests for them. (Indeed, many of the questions on these tests for rationality come straight from K&T experiments!)

But if you look at the average person and ask why they aren't getting what they want, very rarely do you conclude their biggest problem is that they're suffering from anchoring, framing effects, the planning fallacy, commitment bias, or any of the other stuff in the sequences. Usually their biggest problems are far more quotidian and commonsensical.

Take Eliezer. Surely he wanted SIAI to be a well-functioning organization. And he's admitted that lukeprog has done more to achieve that goal of his than he has. Why is lukeprog so much better at getting what Eliezer wants than Eliezer is? It's surely not because lukeprog is so much better at avoiding Sequence-style cognitive biases! lukeprog readily admits that he's constantly learning new rationality techniques from Eliezer.

No, it's because lukeprog did what seems like common sense: he bought a copy of *Nonprofits for Dummies* and did what it recommends. As lukeprog himself says, it wasn't lack of intelligence or resources or akrasia that kept Eliezer from doing these things, "it was a gap in general rationality."

So if you're interested in closing the gap, it seems like the skills to prioritize aren't things like commitment effect and the sunk cost fallacy, but stuff like "figure out what your goals really are", "look at your situation objectively and list the biggest problems", "when you're trying something new and risky, read the *For Dummies* book about it first", etc. For lack of better terminology, let's call the K&T stuff "cognitive biases" and this stuff "practical biases" (even though it's all obviously both practical and cognitive and biases is kind of a negative way of looking at it).

What are the best things you've found on tackling these "practical biases"? Post your suggestions in the comments.

AI timeline predictions: are we getting better?

EDIT: *Thanks to Kaj's work, we now have more rigorous evidence on the "Maes-Garreau law" (the idea that people will predict AI coming before they die). This post has been updated with extra information. The original data used for this analysis can now be found through here.*

Thanks to some sterling work by Kaj Sotala and others (such as Jonathan Wang and Brian Potter - all paid for by the gracious Singularity Institute, a fine organisation that I recommend everyone look into), we've managed to put together a databases listing all AI predictions that we could find. The list is necessarily incomplete, but we found as much as we could, and collated the data so that we could have an overview of what people have been predicting in the field since Turing.

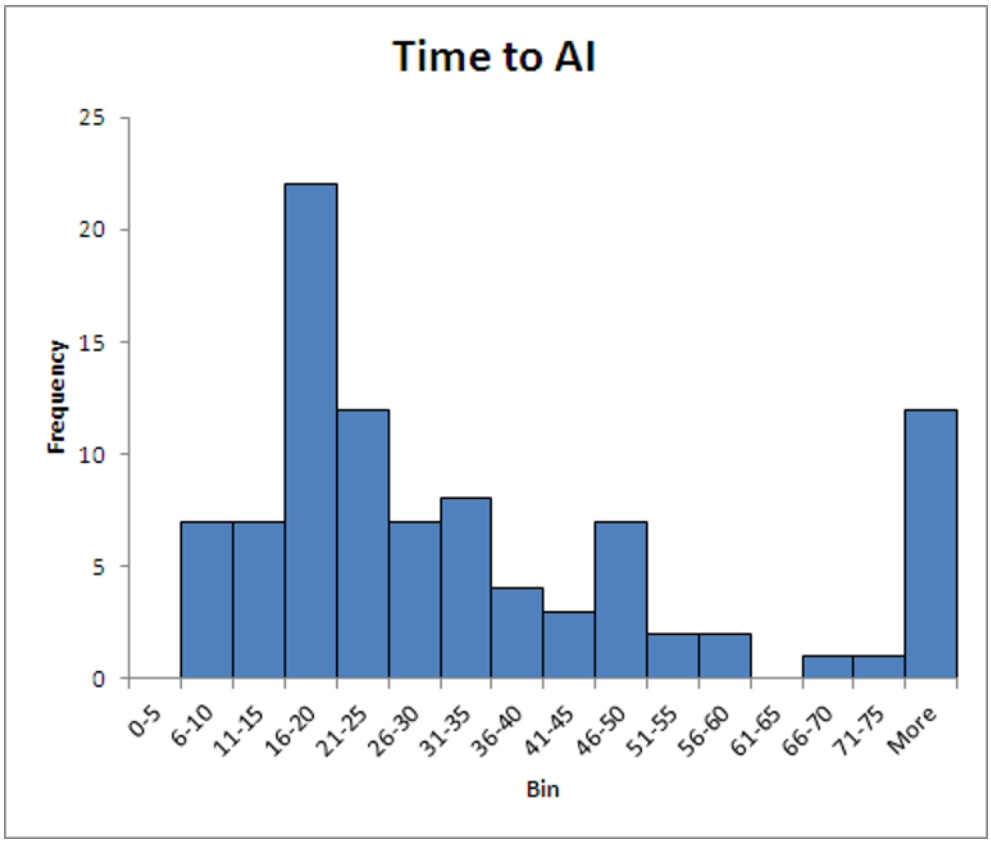
We retained 257 predictions total, of various quality (in our expanded definition, philosophical arguments such as "computers can't think because they don't have bodies" count as predictions). Of these, 95 could be construed as giving timelines for the creation of human-level AIs. And "construed" is the operative word - very few were in a convenient "By golly, I give a 50% chance that we will have human-level AIs by XXXX" format. Some gave ranges; some were surveys of various experts; some predicted other things (such as child-like AIs, or superintelligent AIs).

Where possible, I collapsed these down to single median estimate, making some somewhat arbitrary choices and judgement calls. When a range was given, I took the mid-point of that range. If a year was given with a 50% likelihood estimate, I took that year. If it was the collection of a variety of expert opinions, I took the prediction of the median expert. If the author predicted some sort of AI by a given date (partial AI or superintelligent AI), I took that date as their estimate rather than trying to correct it in one direction or the other (there were roughly the same number of subhuman AIs as suphuman AIs in the list, and not that many of either). I read extracts of the papers to make judgement calls when interpreting problematic statements like "within thirty years" or "during this century" (is that a range or an end-date?).

So some biases will certainly have crept in during the process. That said, it's still probably the best data we have. So keeping all that in mind, let's have a look at what these guys said (and it was mainly guys).

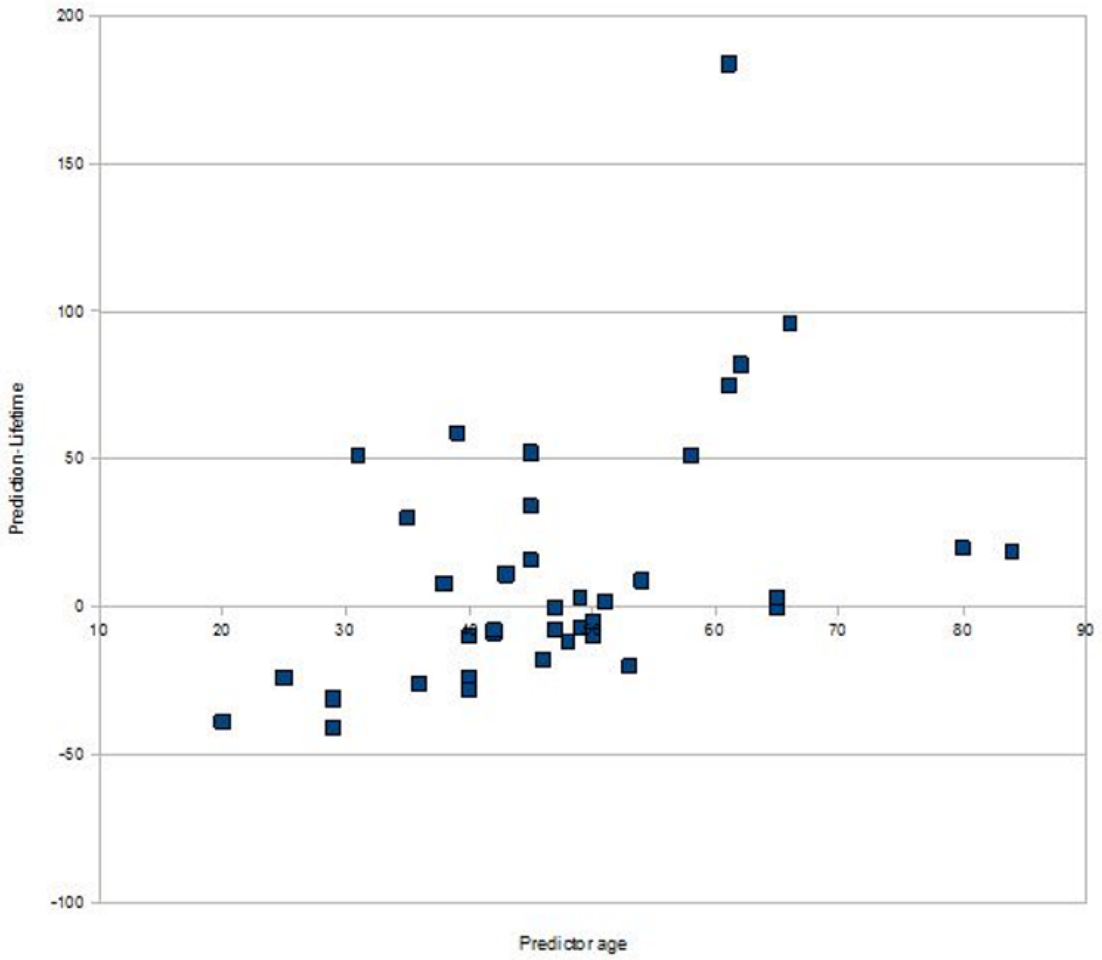
There are two stereotypes about predictions in AI and similar technologies. The first is the Maes-Garreau law: technologies as supposed to arrive... just within the lifetime of the predictor!

The other stereotype is the informal 20-30 year range for any new technology: the predictor knows the technology isn't immediately available, but puts it in a range where people would still be likely to worry about it. And so the predictor gets kudos for addressing the problem or the potential, and is safely retired by the time it (doesn't) come to pass. Are either of these stereotypes born out by the data? Well, here is a histogram of the various "time to AI" predictions:



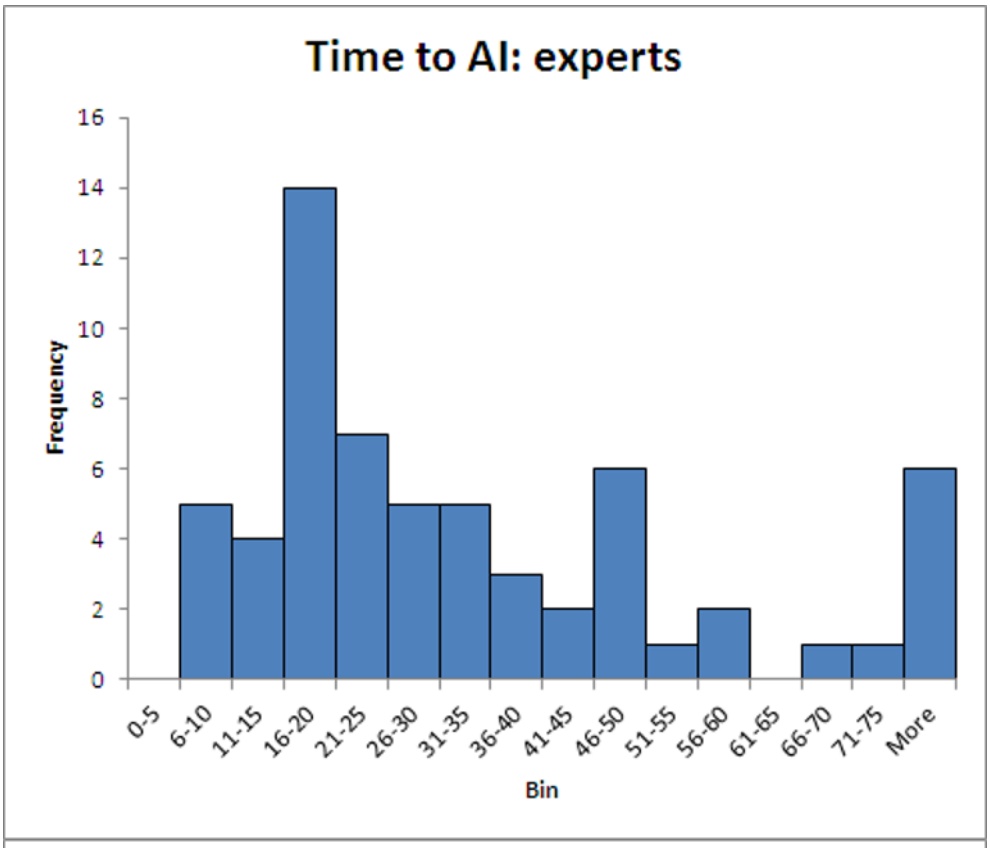
As can be seen, the 20-30 year stereotype is not exactly born out - but a 15-25 one would be. Over a third of predictions are in this range. If we ignore predictions more than 75 years into the future, 40% are in the 15-25 range, and 50% are in the 15-30 range.

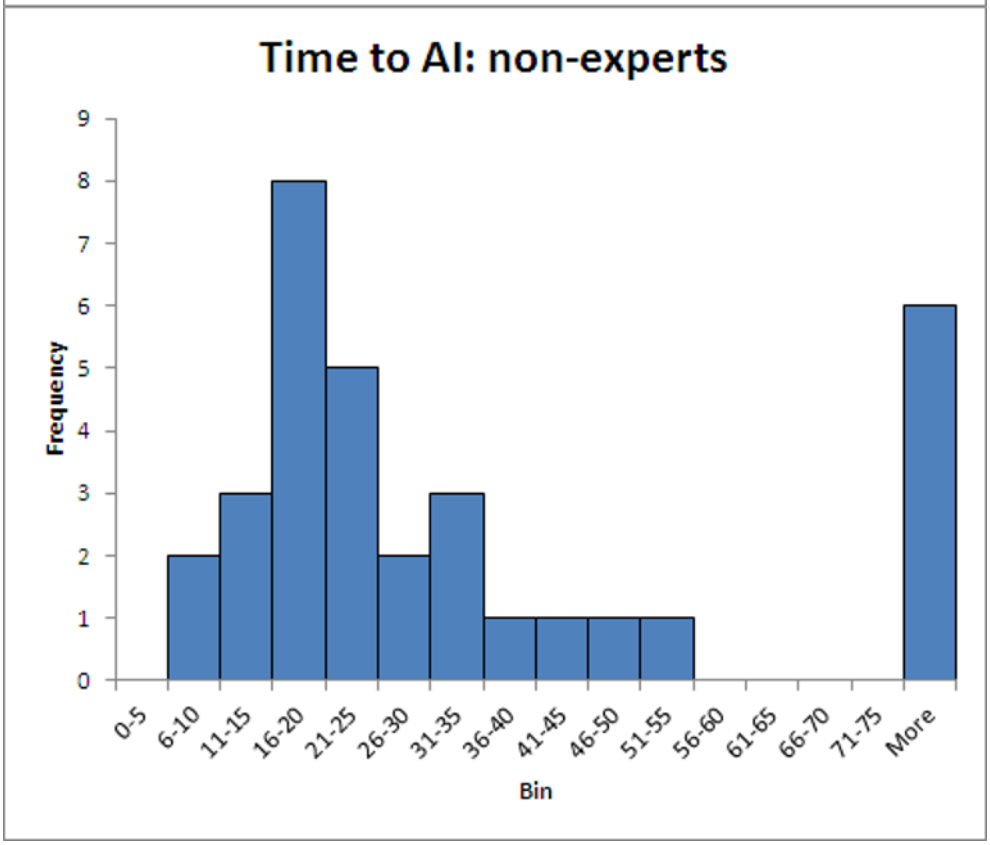
Apart from that, there is a gradual tapering off, a slight increase at 50 years, and twelve predictions beyond three quarters of a century. Eyeballing this, there doesn't seem to much evidence for the Maes-Garreau law. Kaj looked into this specifically, plotting (life expectancy) minus (time to AI) versus the age of the predictor; the Maes-Garreau law would expect the data to be clustered around the zero line:



Most of the data seems to be decades out from the zero point (note the scale on the y axis). You could argue, possibly, that fifty year olds are more likely to predict AI just within their lifetime, but this is a very weak effect. I see no evidence for the Maes-Garreau law - of the 37 prediction Kaj retained, only 6 predictions (16%) were within five years (in either direction) of the expected death date.

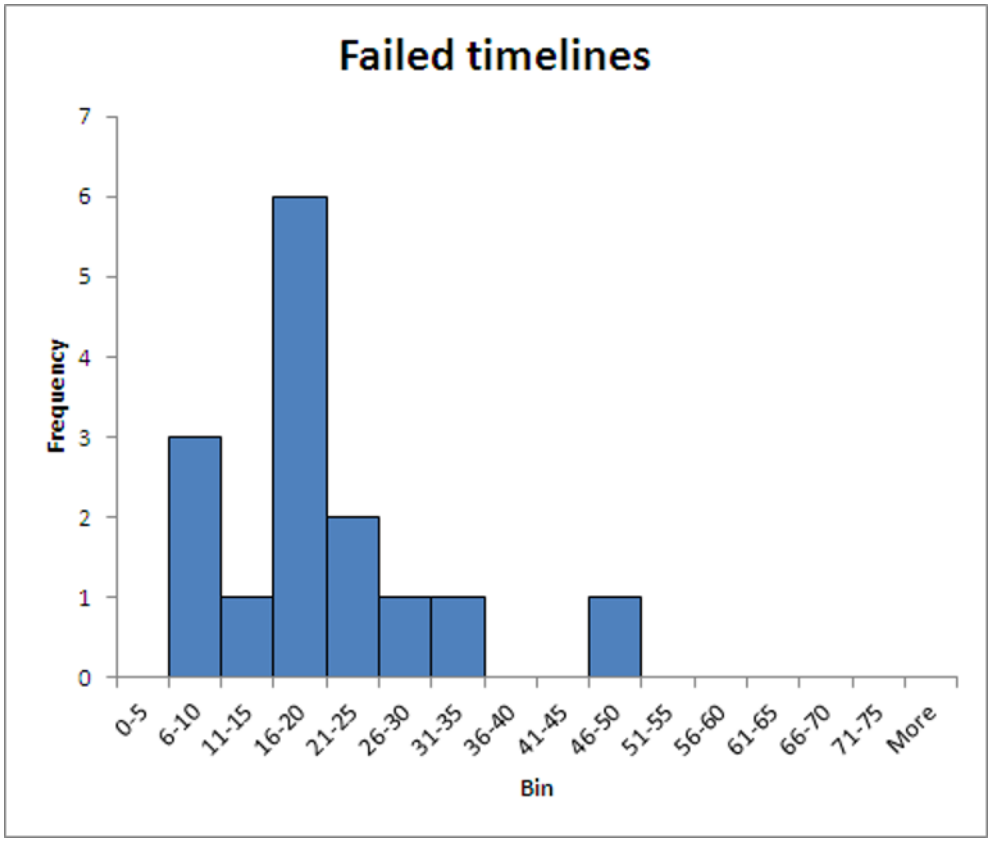
But not all predictions are created equal. 62 of the predictors were labelled "experts" in the analysis - these had some degree of expertise in fields that were relevant to AI. The other 33 were amateurs - journalists, writers and such. Decomposing into these two groups showed very little difference, though:


Time to AI: experts
Frequency
16
14
12
10
8
6
4
2
0
0-5
6-10
11-15
16-20
21-25
26-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70
71-75
More
Bin

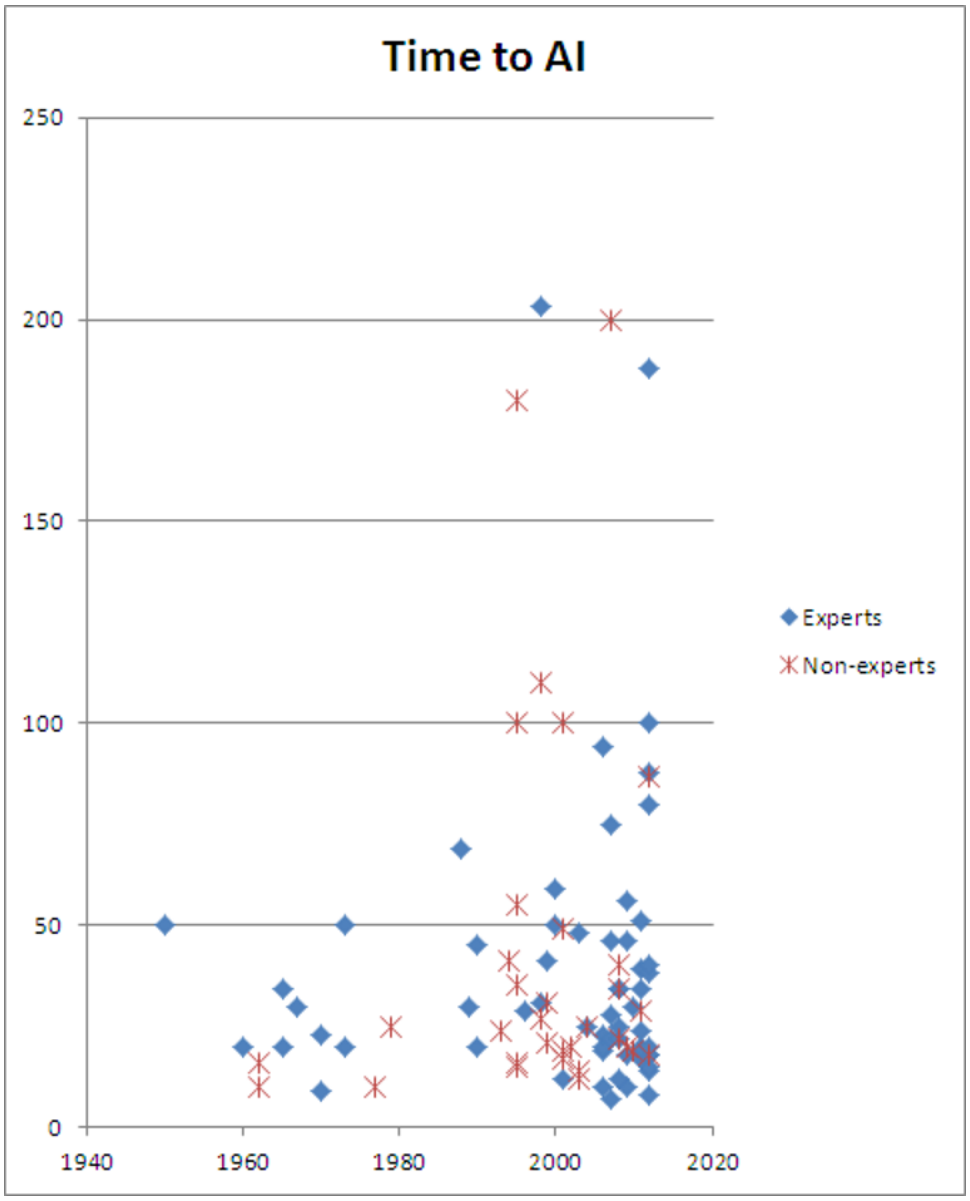

Time to AI: non-experts
Frequency
9
8
7
6
5
4
3
2
1
0
0-5
6-10
11-15
16-20
21-25
26-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70
71-75
More
Bin

The only noticeable difference is that amateurs lacked the upswing at 50 years, and were relatively more likely to push their predictions beyond 75 years. This does not look like good news for the experts - if their performance can't be distinguished from amateurs, what contributions is their expertise making?

But I've been remiss so far - combining predictions that we know are false (because their deadline has come and gone) with those that could still be true. If we look at predictions that have failed, we get this interesting graph:



This looks very similar to the original graph. The main difference being the lack of very long range predictions. This is not, in fact, because there has not yet been enough time for these predictions to be proved false, but because prior to the 1990s, there were actually no predictions with a timeline greater than fifty years. This can best be seen on this scatter plot, which plots the time predicted to AI against the date the prediction was made:



As can be seen, as time elapses, people become more willing to predict very long ranges. But this is something of an artefact - in the early days of computing, people were very willing to predict that AI was impossible. Since this didn't give a timeline, their "predictions" didn't show up on the graph. It recent times, people seem a little less likely to claim AI is impossible, replaced by these "in a century or two" timelines.

Apart from that one difference, predictions look remarkably consistent over the span: modern predictors are claiming about the same time will elapse before AI arrives as their (incorrect) predecessors. This doesn't mean that the modern experts are wrong - maybe AI really is imminent this time round, maybe modern experts have more information and are making more finely calibrated guesses. But in a field like AI prediction, where experts lack feed back for their pronouncements, we should expect them to perform poorly, and for biases to dominate their thinking. This seems the likely hypothesis - it would be extraordinarily unlikely that modern experts, free of biases and full of good information, would reach exactly the same prediction distribution as their biased and incorrect predecessors.

In summary:

- Over a third of predictors claim AI will happen 16-25 years in the future.
- There is no evidence that predictors are predicting AI happening towards the end of their own life expectancy.
- There is little difference between experts and non-experts (some possible reasons for this can be found here).
- There is little difference between current predictions, and those known to have been wrong previously.
- It is not unlikely that recent predictions are suffering from the same biases and errors as their predecessors.

"Epiphany addiction"

LW doesn't seem to have a discussion of the article Epiphany Addiction, by Chris at succeedsocially. First paragraph:

> "Epiphany Addiction" is an informal little term I came up with to describe a process that I've observed happen to people who try to work on their personal issues. How it works is that someone will be trying to solve a problem they have, say a lack of confidence around other people. Somehow they'll come across a piece of advice or a motivational snippet that will make them have an epiphany or a profound realization. This often happens when people are reading self-help materials and they come across something that stands out to them. People can also come up with epiphanies themselves if they're doing a lot of writing and reflecting in an attempt to try and analyze their problems.

I like that article because it describes a dangerous failure mode of smart people. One example was the self-help blog of Phillip Eby (pjeby), where each new post seemed to bring new amazing insights, and after a while you became jaded. An even better, though controversial, example could be Eliezer's Sequences, if you view them as a series of epiphanies about AI research that didn't lead to much tangible progress. (Please don't make that statement the sole focus of discussion!)

The underlying problem seems to be that people get a rush of power from neat-sounding realizations, and mistake that feeling for actual power. I don't know any good remedy for that, but being aware of the problem could help.

Why Don't People Help Others More?

As Peter Singer writes in his book **The Life You Can Save**: "[t]he world would be a much simpler place if one could bring about social change merely by making a logically consistent moral argument". Many people one encounters might agree that a social change movement is noble yet not want to do anything to promote it, or want to give more money to a charity yet refrain from doing so. Additional moralizing doesn't seem to do the trick. ...So what does?

Motivating people to altruism is relevant for the optimal philanthropy movement. For a start on the answer, like many things, I turn to psychology. Specifically, the psychology Peter Singer catalogues in his book.

A Single, Identifiable Victim

One of the most well-known motivations behind helping others is a personal connection, which triggers empathy. When psychologists researching generosity paid participants to join a psychological experiment and then later gave these participants the opportunity to donate to a global poverty fighting organization Save the Children, two different kinds of information were given.

One random group of participants were told "Food shortages in Malawi are affecting more than three million children" and some additional information about how the need for donations was very strong, and these donations could help stop the food shortages.

Another random group of participants were instead shown the photo of Rokia, a seven-year-old Malawian girl who is desperately poor. The participants were told that "her life will be changed for the better by your gift".

Furthermore, a third random group of participants were shown the photo of Rokia, told about who she is and that "her life will be changed for the better", but ALSO told about the general information about the famine and told the same "food shortages [...] are affecting more than three million" -- a combination of both the previous groups.

Lastly, a fourth random group was shown the photo of Rokia, informed about her the same as the other groups, and then given information about another child, identified by name, and told that their donation would also affect this child too for the better.

It's All About the Person

Interestingly, the group who was told ONLY about Rokia gave the most money. The group who was told about both children reported feeling less overal emotion than those who only saw Rokia, and gave less money. The group who was told about both Rokia and the general famine information gave even less than that, followed by the group that only got the general famine information.[1,2] It turns out that information

about a single person was the most salient for creating an empathetic response to trigger a willingness to donate.[1,2]

This continues through additional studies. In another generosity experiment, one group of people was told that a single child needed a lifesaving medical treatment that costs $300K, and was given the opportunity to contribute towards this fund. A second random group of people was told that eight children needed a lifesaving treatment, and all of them would die unless $300K could be provided, and was given an opportunity to contribute. More people opted to donate toward the single child.[3,4]

This is the basis for why we're so willing to chase after lost miners or Baby Jessica no matter the monetary cost, but turn a blind eye to the mass unknown starving in the developing world. Indeed, the person doesn't even need to be particularly identified, though it does help. In another experiment, people asked by researchers to make a donation to Habitat for Humanity were more likely to do so if they were told that the family "has been selected" rather than that they "will be selected" -- even though all other parts of the pitch were the same, and the participants got no information about who the families actually were[5].

The Deliberative and The Affective

Why is this the case? Researcher Paul Slovic thinks that humans have two different processes for deciding what to do. The first is an **affective system** that responds to emotion, rapidly processing images and stories and generating an intuitive feeling that leads to immediate action. The second is a **deliberative system** that draws on reasoning, and operates on words, numbers, and abstractions, which is much slower to generate action.[6]

To follow up, the Rokia experiment was done again, except yet another twist was added -- there were two groups, one told only about Rokia exactly as before, and one told only the generic famine information exactly as before. Within each group, half the group took a survey designed to arouse their emotions by asking them things like "When you hear the word 'baby' how do you feel?" The other half of both groups was given emotionally neutral questions, like math puzzles.

This time, the Rokia group gave far more, but those in the group who randomly had their emotions aroused gave even more than those who heard about Rokia but had finished math problems. On the other side, those who heard the generic famine information showed no increase in donation regardless of how heightened their emotions were.[1]

Futility and Making a Difference

Imagine you're told that there are 3000 refugees at risk in a camp in Rwanda, and you could donate towards aid that would save 1500 of them. Would you do it? And how much would you donate?

Now this time imagine that you can still save 1500 refugees with the same amount of money, but the camp has 10000 refugees. In an experiment where these two scenarios were presented not as a thought experiment but as realities to two separate

random groups, the group that heard of only 3000 refugees were more likely to donate, and donated larger amounts.[7,8]

Enter another quirk of our giving psychology, right or wrong: **futility thinking**. We think that if we're not making a sizable difference, it's not worth making the difference at all -- it will only be a drop in the ocean and the problem will keep raging on.

Am I Responsible?

People are also far less likely to help if they're with other people. In this experiment, students were invited to participate in a market research survey. However, when the researcher gave the students their questionnaire to fill out, she went into a back room separated from the office only by a curtain. A few minutes later, noises strongly suggested that she had got on a chair to get something from a high shelf, and then fell off it, loudly complaining that she couldn't feel or move her foot.

With only one student taking the survey, 70% of them stopped what they were doing and offered assistance. However, when there were two students taking the survey, this number dropped down dramatically. Most noticeably, when the group was two students -- but one of the students was a stooge who was in on it and would always not respond, the response rate of the non-stooge participant was only 7%.[9]

This one is known as **diffusion of responsibility**, better known as the **bystander effect** -- we help more often when we think it is our responsibility to do so, and -- again for right or for wrong -- we naturally look to others to see if they're helping before doing so ourselves.

What's Fair In Help?

It's clear that people value fairness, even to their own detriment. In a game called "the Ultimatum Game", one participant is given a sum of money by the researcher, say $10, and told they can split this money with an anonymous second player in any proportion they choose -- give them $10, give them $7, give them $5, give them nothing, everything is fair game. The catch is, however, the second player, after hearing of the split anonymously, gets to vote to accept it or reject it. Should the split be accepted, both players walk away with the agreed amount. But should the split be rejected, both players walk away with nothing.

A Fair Split

The economist, expecting ideally rational and perfectly self-interested players, predicts that the second player would accept any split that gets them money, since anything is better than nothing. And the first player, understanding this, would naturally offer $1 and keep $9 for himself. At no point are identities revealed, so reputation and retribution are no issue.

But the results turn out to be quite different -- the vast majority offer an equal split. Yet, when an offer comes around that offers $2 or less, it is almost always rejected, even though $2 is better than nothing.[10] And this effect persists even when played for thousands of dollars and persists across nearly all cultures.

Splitting and Anchoring in Charity

This sense of fairness persists into helping as well -- people generally have a strong tendency not to want to help more than the other people around them, and if they find themselves the only ones helping on a frequent basis, they start to feel a "sucker". On the flipside, if others are doing more, they will follow suit.[11,12,13]

Those told the average donation to a charity nearly always tend to give that amount, even if the average told to them is a lie, having secretly been increased or decreased. And it can be replicated even without lying -- those told about an above average gift were far more likely to donate more, even attempting to match that gift.[14,15] Overall, we tend to match the behavior of our reference class -- those people we identify with -- and this includes how much we help. We donate more when we believe others are donating more, and donate less when we believe others are doing so.

Challenging the Self-Interest Norm

But there's a way to break this cycle of futility, responsibility, and fairness -- challenge the norm by openly communicating about helping others. While many religious and secular values insist that the best giving is anonymous giving, this turns out to not always be the case. While there may be other reasons to give anonymously, don't forget the benefits of giving openly -- being open about helping inspires others to help, and can help challenge the norms of the culture.

Indeed, many organizations now exist to help challenge the norms of donations and try to create a culture where they give more. GivingWhatWeCan is a community of 230 people (including me!) who have all pledged to donate at least 10% of their income to organizations working on ending extreme poverty, and submit statements proving so. BolderGiving has a bunch of inspiring stories of over 100 people who all give at least 20% of their income, with a dozen giving over 90%! And these aren't all rich people, some of them are even ordinary students.

Who's Willing to Be Altruistic?

While people are not saints, experiments have shown that people tend to grossly overestimate how self-interested other people are -- for one example, people estimated that males would overwhelmingly favor a piece of legislation to "slash research funding to a disease that affects only women", even while -- being male -- they themselves do not support such legislation.[16]

This also manifests itself in an expectation that people be "self-interested" in their philanthropic cause -- suggesting much stronger support for volunteers in Students Against Drunk Driving who themselves knew people killed in drunk driving accidents

versus those people who had no such personal experiences but just thought it to be "a very important cause".[17]

Alex de Tocqueville, echoing the early economists who expected $9/$1 splits in the Ultimatum Game, wrote in 1835 that "Americans enjoy explaining almost every act of their lives on the principle of self-interest".[18] But this isn't always the case, and in challenging the norm, people make it more acceptable to be altruistic. It's not just "goody two-shoes", and it's praiseworthy to be "too charitable".

A Bit of a Nudge

A somewhat pressing problem in getting people to help was in organ donation -- surely no one was inconvenienced by having their organs donated after they had died. Yet, why would people not sign up? And how could we get more people to sign up?

In Germany, only 12% of the population are registered organ donors. In nearby Austria, that number is 99.98%. Are people in Austria just less worried about what will happen to them after they die, or just that more altruistic? It turns out the answer is far more simple -- in Germany you must put yourself on the register to become a potential donor (opt-in), whereas in Austria you are a potential donor unless you object (opt-out). While people may be, for right or for wrong, worried about the fate of their body after it is dead, they appear less likely to express these reservations in opt-out systems.[19]

While Richard Thaler and Cass Sunstein argue in their book **Nudge: Improving Decisions About Health, Wellness, and Happiness** that we sometimes suck at making decisions in our own interest and all could do better with more favorable "defaults", such defaults are also pressing in helping people.

While opt-out organ donation is a huge deal, there's another similar idea -- **opt-out philanthropy**. Back before 2008 when the investment bank Bear Stearns still existed, Bear Stearns listed their guiding principle as philanthropy as fostering good citizenship and well-rounded individuals. To this effect, they required the top 1000 most highest paid employees to donate 4% of their salary and bonuses to non-profits, and prove it with their tax returns. This resulted in more than $45 million in donations during 2006. Many employees described the requirement as "getting themselves to do what they wanted to do anyway".

Conclusions

So, according to this bit of psychology, what could we do to get other people to help more, besides moralize? Well, we have five key take-aways:

(1) present these people with a single and highly identifiable victim that they can help
(2) nudge them with a default of opt-out philanthropy
(3) be more open about our willingness to be altruistic and encourage other people to help

(4) make sure people understand the average level of helping around them, and (5) instill a responsibility to help and an understanding that doing so is not futile.

Hopefully, with these tips and more, helping people more can be come just one of those things we do.

References

(Note: Links are to PDF files.)

1: D. A. Small, G. Loewenstein, and P. Slovic. 2007. "Sympathy and Callousness: The Impact of Deliberative Thought on Donations to Identifiable and Statistical Victims". *Organizational Behavior and Human Decision Processes* 102: p143-53

2: Paul Slovic. 2007. **"If I Look at the Mass I Will Never Act: Psychic Numbing and Genocide"**. *Judgment and Decision Making* 2(2): p79-95.

3: T. Kogut and I. Ritov. 2005. "The 'Identified Victim' Effect: An Identified Group, or Just a Single Individual?". *Journal of Behavioral Decision Making* 18: p157-67.

4: T. Kogut and I. Ritov. 2005. "The Singularity of Identified Victims in Separate and Joint Evaluations". *Organizational Behavior and Human Decision Processes* 97: p106-116.

5: D. A. Small and G. Lowenstein. 2003. "Helping the Victim or Helping a Victim: Altruism and Identifiability". *Journal of Risk and Uncertainty* 26(1): p5-16.

6: Singer cites this from Paul Slovic, who in turn cites it from: Seymour Epstein. 1994. "Integration of the Cognitive and the Psychodynamic Unconscious". *American Psychologist* 49: p709-24. Slovic refers to the affective system as "experiential" and the deliberative system as "analytic". This is also related to Daniel Kahneman's popular book Thinking Fast and Slow.

7: D. Fetherstonhaugh, P. Slovic, S. M. Johnson, and J. Friedrich. 1997. "Insensitivity to the Value of Human Life: A Study of Psychophysical Numbing". *Journal of Risk and Uncertainty* 14: p283-300.

8: Daniel Kahneman and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica* 47: p263-91.

9: Bib Lantané and John Darley. 1970. The Unresponsive Bystander: Why Doesn't He Help?. New York: Appleton-Century-Crofts, p58.

10: Martin Nowak, Karen Page, and Karl Sigmund. 2000. "Fairness Versus Reason in the Ultimatum Game". *Science* 289: p1183-75.

11: Lee Ross and Richard E. Nisbett. 1991. The Person and the Situation: Perspectives of Social Psychology. Philadelphia: Temple University Press, p27-46.

12: Robert Cialdini. 2001. Influence: Science and Practice, 4th Edition. Boston: Allyn and Bacon.

13: Judith Lichtenberg. 2004. "Absence and the Unfond Heart: Why People Are Less Giving Than They Might Be". in Deen Chatterjee, ed. The Ethics of Assistance: Morality and the Distant Needy. Cambridge, UK: Cambridge University Press.

14: Jen Shang and Rachel Croson. Forthcoming. "Field Experiments in Charitable Contribution: The Impact of Social Influence on the Voluntary Provision of Public Goods". *The Economic Journal*.

15: Rachel Croson and Jen Shang. 2008. "The Impact of Downward Social Information on Contribution Decision". *Experimental Economics* 11: p221-33.

16: Dale Miller. 199. "The Norm of Self-Interest". *American Psychologist* 54: 1053-60.

17: Rebecca Ratner and Jennifer Clarke. Unpublished. "Negativity Conveyed to Social Actors Who Lack a Personal Connection to the Cause".

18: Alexis de Tocqueville in J.P. Mayer ed., G. Lawrence, trans. 1969. Democracy in America. Garden City, N.Y.: Anchor, p546.

19: Eric Johnson and Daniel Goldstein. 2003. "Do Defaults Save Lives?". Science 302: p1338-39.

(This is an updated version of an earlier draft from my blog.)

Solving the two envelopes problem

Suppose you are presented with a game. You are given a red and a blue envelope with some money in each. You are allowed to ask an independent party to open both envelopes, and tell you the ratio of blue:red amounts (but not the actual amounts). If you do, the game master replaces the envelopes, and the amounts inside are chosen by him using the same algorithm as before.

You ask the independent observer to check the amounts a million times, and find that half the time the ratio is 2 (blue has twice as much as red), and half the time it's 0.5 (red has twice as much as blue). At this point, the game master discloses that in fact, the way he chooses the amounts mathematically guarantees that these probabilities hold.

Which envelope should you pick to maximize your expected wealth?

It may seem surprising, but with this set-up, the game master can choose to make either red or blue have a higher expected amount of money in it, or make the two the same. Asking the independent party as described above will not help you establish which is which. This is the surprising part and is, in my opinion, the crux of the two envelopes problem.

This is not quite how the two envelopes problem is usually presented, but this is the presentation I arrived at after contemplating the original puzzle. The original puzzle prescribes a specific strategy that the game master follows, makes the envelopes indistinguishable, and provides a paradoxical argument which is obviously false, but it's not so obvious where it goes wrong.

Note that for simplicity, let's assume that money is a real quantity and can be subdivided indefinitely. This avoids the problem of odd amounts like $1.03 not being exactly divisible by two.

The flawed argument

The flawed argument goes as follows. Let's call the amount in the blue envelope B, and in red R. You have confirmed that half the time, B is equal to 2R, and half the time it's R/2. This is a fact. Surely then the expected value of B is (2R * 50% + R/2 * 50%), which simplifies to 1.25R. In other words, the blue envelope has a higher expected amount of money given the evidence we have.

But notice that the situation is completely symmetric. From the information you have, it's also obvious that half the time, R is 2B, and half the time it's B/2. So by the same argument the expected value of R is 1.25B. Uh-oh. The expected value of both envelopes is higher than the other?...

Game master strategies

Let's muddy up the water a little by considering the strategies the game master can use to pick the amounts for each envelope.

Strategy 1

Pick an amount X between $1 and $1000 randomly. Throw a fair die. If you get an odd number, put X into the red envelope and 2X into blue. Otherwise put X into blue and 2X into red.

Strategy 2

Pick an amount X between $1 and $1000 randomly. Put this into the red envelope. Throw a fair die. If you get an odd number, put 2X into blue, and if it's even, put X/2 into blue.

The difference between these strategies is fairly subtle. I hope it's sufficiently obvious that the "*ratio condition*" (B = 2R half the time and R = 2B the other half) is true for both strategies. However, suppose we have two people take part in this game, one always picking the red envelope and the other always picking the blue envelope. After a million repetitions of this game, with the first strategy, the two guys will have won almost exactly the same amounts in total. After a million repetitions with the second strategy, the total amount won by the blue guy will be *25% higher* than the total amount won by the red guy!

Now observe that strategy 2 can be trivially inverted to favour the red envelope instead of the blue one. The player can ask an independent observer for ratios (as described in the introduction) all he wants, but this information will not allow him to distinguish between these three scenarios (strategy 1, strategy 2 and strategy 2 inverted). It's obviously impossible to figure out which envelope has a higher expected winnings from this information!

What's going on here?

I hope I've convinced you by now that the information about the likelihood of the ratios does not tell you which envelope is better. But what *exactly* is the flaw in the original argument?

Let's formalize the puzzle a bit. We have two random variables, R and B. We are permitted to ask someone to sample each one and compute the ratio of the samples, r/b, and disclose it to us. Let's define a random variable called RB whose samples are produced by sampling R and B and computing their ratio. We know that RB can take two values, 2 and 0.5, with equal probability. Let's also define BR, which is the opposite ratio: that of a sample of B to a sample of R. BR can also take two values, 2 and 0.5, with equal probability.

The flawed argument is simply that the expected value of RB, E(RB), is 1.25, which is greater than 1, and therefore E(R) > E(B). The flawed argument continues that E(BR) is 1.25 too, therefore E(B) > E(R), leading to a contradiction. What's the flaw?

Solution

The expected value of RB, E(RB), really is 1.25. The puzzle gets that one right. E(BR) is *also* 1.25. The flaw in the argument is simply that it assumes E(X/Y) > 1 implies that E(X) > E(Y). This implication seems to hold intuitively, but human intuition is notoriously bad at probabilities. It is easy to prove that this implication is false, by considering a simple counter-example courtesy of VincentYu.

Consider two independent random variables, X and Y. X can take values 20 and 60, while Y can take values 2 and 100, both with equal probability. To calculate the expected value of X/Y, one can enumerate all possible combinations, multiplying each by its probability. The four possible combinations of X and Y are 20/2, 20/100, 60/2 and 60/100. Each combination is 25% likely. Hence E(X/Y) is 10.2. This is greater than 1, so the if the implication were to hold, E(X) should be greater than E(Y). But E(X) is (20+60)/2 = 40, while E(Y) is (2+100)/2 = 51. Hence, the implication E(X/Y) > 1 => E(X) > E(Y) does not hold in general.

So there you have it. The proposed argument relies on an implication which seems true intuitively, but turns out to be false under scrutiny. Mystery solved?... Almost.

Imprecise language's contribution to the puzzle

The argument concerning the original, indistinguishable envelopes, is phrased like this: "*(1) I denote by A the amount in my selected envelope. (2) The other envelope may contain either 2A or A/2, with a 50% probability each. (3) So the expected value of the money in the other envelope is 1.25A. (4) Hence, the other envelope is expected to have more dollars.*"

Depending on how pedantic you are, you might say that the statement made in the third sentence is strictly false, or that it is too ambiguous to be strictly false, or that at least one interpretation is true. The expected value 1.25A is "*of the amount of money contained in the other envelope expressed in terms of the amount of money in this envelope*". It is ***not*** "*of the amount of money in the other envelope expressed in dollars*". Hence the last sentence does not follow, and if the statements were made in full and with complete accuracy, the fact that it does not follow is a little bit more obvious.

In closing, I would say this puzzle is hard because "in terms of this envelope" and "in terms of dollars" are typically equivalent enough in everyday life, but when it comes to expected values, this equivalence breaks down rather counter-intuitively.

Who Wants To Start An Important Startup?

SUMMARY: *Let's collect people who want to work on for-profit companies that have significant positive impacts on many people's lives.*

Google provides a huge service to the world - efficient search of a vast amount of data. I would really like to see more for-profit businesses like Google, especially in underserved areas like those explored by non-profits GiveWell, Singularity Institute and CFAR. GiveWell is a nonprofit that is both working toward making humanity better, and thinking about leverage. Instead of hacking away at one branch of the problem of effective charity by working on one avenue for helping people, they've taken it meta. They're providing a huge service by helping people choose non-profits to donate to that give the most bang for your buck, and they're giving the non-profits feedback on how they can improve. I would love to see more problems taken meta like that, where people invest in high leverage things.

Beyond these non-profits, I think there is a huge amount of low-hanging fruit for creating businesses that create a lot of good for humanity and make money. For-profit businesses that pay their employees and investors well have the advantage that they can entice very successful and comfortable people away from other jobs that are less beneficial to humanity. Unlike non-profits where people are often trying to scrape by, doing the good of their hearts, people doing for-profits can live easy lives with luxurious self care while improving the world at the same time.

It's all well and good to appeal to altruistic motives, but a lot more people can be mobilzed if they don't have to sacrifice their own comfort. I have learned a great deal about this from Jesse and Sharla at Rejuvenate. They train coaches and holistic practitioners in sales and marketing - enabling thousands of people to start businesses who are doing the sorts of things that advance their mission. They do this while also being multi-millionaires themselves, and maintaining a very comfortable lifestyle, taking the time for self-care and relaxation to recharge from long workdays.

Less Wrong is read by thousands of people, many of whom are brilliant and talented. In addition, Less Wrong readers include people who are interested in the future of the world and think about the big picture. They think about things like AI and the vast positive and negative consequences it could have. In general, they consider possibilities that are outside of their immediate sensory experience.

I've run into a lot of people in this community with some really cool, unique, and interesting ideas, for high-impact ways to improve the world. I've also run into a lot of talent in this community, and I have concluded that we have the resources to implement a lot of these same ideas.

Thus, I am opening up this post as a discussion for these possibilities. I believe that we can share and refine them on this blog, and that there are talented people who will execute them if we come up with something good. For instance, I have run into countless programmers who would love to be working on something more inspiring than what they're doing now. I've also personally talked to several smart organizational leader types, such as Jolly and Evelyn, who are interested in helping with and/or leading inspiring projects And that's only the people I've met personally; I

know there are a lot more folks like that, and people with talents and resources that haven't even occurred to me, who are going to be reading this.

Topics to consider when examining an idea:

- Tradeoffs between optimizing for good effects on the world v. making a profit.
- Ways to improve both profitability and good effects on the world.
- Timespan - projects for 3 months, 1 year, 5 years, 10+ years
- Using resources efficiently (e.g. creating betting markets where a lot of people give opinions that they have enough confidence in to back with money, instead of having one individual trying to figure out probabilities)
- Opportunities for uber-programmers who can do anything quickly (they are reading and you just might interest and inspire them)
- Opportunities for newbies trying to get a foot in the door who will work for cheap
- What people/resources do we have at our disposal now, and what can we do with that?
- What people/resources are still needed?
- If you think of something else, make a comment about it in the thread for that, and it might get added to this list.

An example idea from Reichart Von Wolfsheild:

> *A project to document the best advice we can muster into a single tome. It would inherently be something dynamic, that would grow and cover the topics important to humans that they normally seek refuge and comfort for in religion. A "bible" of sorts for the critical mind.*
>
> *Before things like wikis, this was a difficult problem to take on. But, that has changed, and the best information we have available can in fact be filtered for, and simplified. The trick now, is to organize it in a way that helps humans. which is not how most information is organized.*

Collaboration

1. **Please keep the mission in mind** (let's have more for-profit companies working on goals that benefit people too!) when giving feedback. When you write a comment, consider whether it is contributing to that goal, or if it's counterproductive to motivation or idea-generation, and edit accordingly.
2. **Give feedback, the more specific the better.** Negative feedback is valuable because it tells us where to concentrate further work. It can also be a motivation-killer; it feels like punishment, and not just for the specific item criticized, so be charitable about the motives and intelligence of others, and stay mindful of how much and how aggressively you dole critiques out. (Do give critiques, they're essential - just be gentle!) Also, distribute positive feedback for the opposite effect. More detail on giving the best possible feedback in this comment.
3. **Please point other people with resources such as business experience, intelligence, implementation skills, and funding capacity at this post.** The more people with these resources who look at this and collaborate in the comments, the more likely it is for these ideas to get implemented. In addition to posting this to Less Wrong, I will be sending the link to a lot of friends with shrewd business skills, resources and talent, who might be interested in helping

make projects happen, or possibly in finding people to work on their own projects since many of them are already working on projects to make the world better.

4. **Please provide feedback**. If anything good happens in your life as a result of this post or discussion, please comment about it and/or give me feedback. It inspires people, and I have bets going that I'd like to win. Consider making bets of your own! It is also important to let me know if you are going to use the ideas, so that we don't end up with needless duplication and competition.

Finally: If this works right, there will be lots of information flying around. Check out the organization thread and the wiki.

Self-skepticism: the first principle of rationality

When Richard Feynman started investigating irrationality in the 1970s, he quickly begun to realize the problem wasn't limited to the obvious irrationalists.

Uri Geller claimed he could bend keys with his mind. But was he really any different from the academics who insisted their special techniques could teach children to read? Both failed the crucial scientific test of skeptical experiment: Geller's keys failed to bend in Feynman's hands; outside tests showed the new techniques only caused reading scores to go down.

What mattered was not how smart the people were, or whether they wore lab coats or used long words, but whether they followed what he concluded was the crucial principle of truly scientific thought: "a kind of utter honesty--a kind of leaning over backwards" to prove yourself wrong. In a word: self-skepticism.

As Feynman wrote, "The first principle is that you must not fool yourself -- and you are the easiest person to fool." Our beliefs always seem correct to us -- after all, that's why they're our beliefs -- so we have to work extra-hard to try to prove them wrong. This means constantly looking for ways to test them against reality and to think of reasons our tests might be insufficient.

When I think of the most rational people I know, it's this quality of theirs that's most pronounced. They are constantly trying to prove themselves wrong -- they attack their beliefs with everything they can find and when they run out of weapons they go out and search for more. The result is that by the time I come around, they not only acknowledge all my criticisms but propose several more I hadn't even thought of.

And when I think of the least rational people I know, what's striking is how they do the exact opposite: instead of viciously attacking their beliefs, they try desperately to defend them. They too have responses to all my critiques, but instead of acknowledging and agreeing, they viciously attack my critique so it never touches their precious belief.

Since these two can be hard to distinguish, it's best to look at some examples. The Cochrane Collaboration argues that support from hospital nurses may be helpful in getting people to quit smoking. How do they know that? you might ask. Well, they found this was the result from doing a meta-analysis of 31 different studies. But maybe they chose a biased selection of studies? Well, they systematically searched "MEDLINE, EMBASE and PsycINFO [along with] hand searching of specialist journals, conference proceedings, and reference lists of previous trials and overviews." But did the studies they pick suffer from selection bias? Well, they searched for that -- along with three other kinds of systematic bias. And so on. But even after all this careful work, they still only are confident enough to conclude "the results…support a modest but positive effect…with caution … these meta-analysis findings need to be interpreted carefully in light of the methodological limitations".

Compare this to the Heritage Foundation's argument for the bipartisan Wyden–Ryan premium support plan. Their report also discusses lots of objections to the proposal, but confidently knocks down each one: "this analysis relies on two highly implausible assumptions … All these predictions were dead wrong. … this perspective completely

ignores the history of Medicare" Their conclusion is similarly confident: "The arguments used by opponents of premium support are weak and flawed." Apparently there's just not a single reason to be cautious about their enormous government policy proposal!

Now, of course, the Cochrane authors might be secretly quite confident and the Heritage Foundation might be wringing their hands with self-skepticism behind-the-scenes. But let's imagine for a moment that these aren't just reportes intended to persuade *others* of a belief and instead accurate portrayals of how these two different groups approached the question. Now ask: which style of thinking is more likely to lead *the authors* to the right answer? Which attitude seems more like Richard Feynman? Which seems more like Uri Geller?

A model of UDT with a concrete prior over logical statements

I've been having difficulties with constructing a toy scenario for AI self-modification more interesting than Quirrell's game, because you really want to do expected utility maximization of some sort, but currently our best-specified decision theories search through the theorems of one particular proof system and "break down and cry" if they can't find one that tells them what their utility will be if they choose a particular option. This is fine if the problems are simple enough that we always find the theorems we need, but the AI rewrite problem is precisely about skirting that edge. It seems natural to want to choose some probability distribution over the possibilities that you can't rule out, and then do expected utility maximization (because if you don't maximize EU over some prior, it seems likely that someone could Dutch-book you); indeed, Wei Dai's original UDT has a "mathematical intuition module" black box which this would be an implementation of. But how *do* you assign probabilities to logical statements? What consistency conditions do you ask for? What are the "impossible possible worlds" that make up your probability space?

Recently, Wei Dai suggested that logical uncertainty might help avoid the Löbian problems with AI self-modification, and although I'm sceptical about this idea, the discussion pushed me into trying to confront the logical uncertainty problem head-on; then, reading Haim Gaifman's paper "Reasoning with limited resources and assigning probabilities to logical statements" (which Luke linked from So you want to save the world) made something click. I want to present a simple suggestion for a concrete definition of "impossible possible world", for a prior over them, and for an UDT algorithm based on that. I'm not sure whether the concrete prior is useful—the main point in giving it is to have a concrete example we can try to prove things about—but the definition of logical possible worlds looks like a promising theoretical tool to me.

*

Let S be the set of sentences of Peano Arithmetic less than 3^^^3 symbols long, and let Pow(S) := the set of all subsets of S. We interpret elements X of Pow(S) as "logical worlds", in which the sentences in X are true and the sentences in (S \ X) are false. Each world X can be represented by a single sentence s_X, which is the conjunction of the sentences ({x | x in X} union {not y | y in S \ X}). Now, we exclude those X that are proven contradictory by one of the first 4^^^^4 theorems of PA; that is, we define the set W of *possible worlds* to be

W := {X in Pow(S) | "not s_X" is **not** among the first 4^^^^4 theorems of PA}.

W is our probability space. Note that it's finite. For lack of a better idea, I propose to use the uniform prior on it. (Possibly useful: another way to think about this is that we choose the uniform prior on Pow(S), and after looking at the 4^^^^4 theorems, we do a Bayesian update.)

Individual sentences in S induce events over this probability space: a sentence x corresponds to the event {X in Pow(S) | x in X}. Clearly, all of the above can be carried out by a computable function, and in particular we can write a computable function P(.) which takes a statement in S and returns the probability of the

corresponding event (and the source of this function can be short, i.e., it doesn't need to contain any 3^^^3-sized lookup tables).

*

The decision algorithm makes use of two global variables which specify the problem.

- `actions` is a Python dictionary that maps possible inputs to the algorithm to a list of outputs the algorithm is allowed to return when receiving that input. For example, in the problem from Wei Dai's UDT1.1 post, `actions = {1: ['A','B'], 2: ['A','B']}`, meaning that your input is either '1' or '2', and in both cases you may choose between options 'A' and 'B'. We'll assume there's nothing 3^^^3-sized in `actions`.
- `worlds` is a list of triples of the form `(p,U,us)`, representing possible physical worlds the agent might find itself in, where `p` is the probability of being in that world, `U` is the source of a function that computes and returns the agent's utility if that world is the true one (by simulating that world and running a utility function over the result), and `us` is a list of values `U` might return. The probabilities must add to 1. We'll assume that there's nothing 3^^^3-sized here either, and that it's provable in much less that 4^^^^4 steps that *if* the decision algorithm halts on all inputs specified by `actions` and returns one of the allowable actions, then each `U` will halt and return a value in the corresponding `us`. (The reason for the condition is that the functions `U` may contain copies of the agent's source, and may make calls to the agent, so if the agent didn't halt, neither could `U`.)

With these provisions, the algorithm, `UDT(input)`, proceeds as follows:

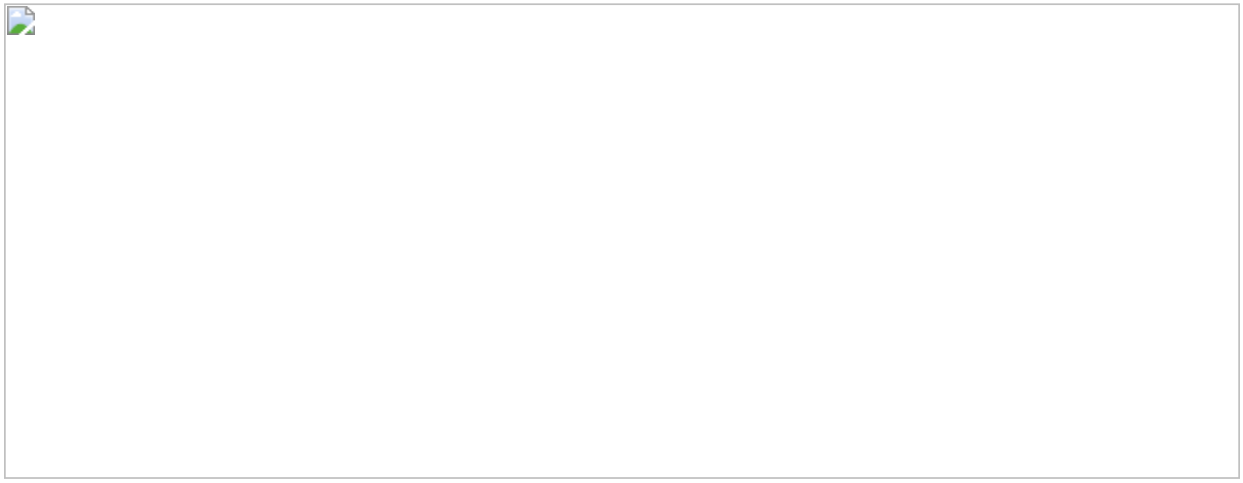
1. Compute `mappings`, a list of all dictionaries that maps each possible input from `actions` to one of the allowable outputs. (In the earlier example, `mappings = [{1:'A',2:'A'}, {1:'A',2:'B'}, {1:'B',2:'A'}, {1:'B',2:'B'}]`.)
2. Play chicken with the universe: For each `m` in `mappings`, if P("`UDT(i) == m[i]` for every `i` in `actions.keys()`") = 0, then return `m[input]`.
3. Calculate expected utilities: For each `m` in `mappings`, for each `(p,U,us)` in `worlds`, for each `u` in `us`, compute `q` := P("`U() == u`" | "`UDT(i) == m[i]` for every `i` in `actions.keys()`"); the expected utility EU(`m`) of `m` is the sum of all the corresponding `p*u*q`. (Note that we made sure in the previous step that the conditional probabilities `q` exist.)
4. Choose the `m` with the highest expected utility. If multiple options have the same utility, choose the lexicographically lowest one.
5. Return `m[input]`.

Now, the universe must always chicken out (the algorithm will never need to return in step 2), because one of the possible worlds in W must be *true*, this true world cannot be ruled out by Peano Arithmetic because PA is sound, and if the algorithm returned `m[input]` in step 2, then "`UDT(i) == m[i]` for every `i` in `actions.keys()`" would hold in this true world, so the probability of this sentence could not be zero.

Further, the algorithm will halt on all inputs, because although it does some big computations, there is no *unbounded* search anywhere; and it's easy to see that on each possible input, it will return one of the allowable outputs. This reasoning can be formalized in PA. Using our earlier assumption, it follows (in PA, in much less than 4^^^^4 steps) that each `U` will halt and return a value in the corresponding `us`. Thus, in each possible logical world in W, for every `(p,U,us)` in `worlds`, "`U()` halts" will be true,

and "`U() == u`" will be true for exactly one `u` (more than one would quickly prove a contradiction), and this `u` will be in `us`; and therefore, the different `q` for a given `(p,U,us)` will add up to 1.

The High Impact Network (THINK) - Launching Now



THINK, The High Impact Network, is going live this week.

We're a network of Effective Altruists (EAs), looking to do the most good for the most people[1] as efficiently as possible. We aren't bound by a central cause or ethical framework, but rather by a process, and a commitment to rigor and rationality as we try to make the world a better place.

THINK meetups are forming around the world. Some are functioning as student groups at prominent universities, others are general meetups for people of all ages who want to make effective altruism a part of their life. As I write this, 20 meetups are getting ready to launch in the fall, and discussions are underway for an additional 30. If you'd like to connect with other EA-types, see if a meetup's forming in your area, or run your own meetup, send us an e-mail here, or visit our website.

We're putting together a collection of meetup modules, which newly formed groups can use for content at weekly meetups. These fall into roughly two categories:

- Introductory materials, designed to teach the basics of Effective Altruism to newcomers.
- Self Improvement tools, helping newcomers and veterans to become strong enough to tackle the difficult problems ahead.

Five sample modules are available on our website, and more are coming. If you have ideas for a module and would like to create you own, e-mail us at modules@thehighimpactnetwork.org.

But most importantly - we want bright, enthusiastic people who care deeply about the world to collaborate with each other on high impact projects.

Optimal Philanthropy. Effective Altruism.

Less Wrong veterans will recognize the basics of Optimal Philanthropy, although we consider avenues beyond traditional charity. (The phrase "effective altruism" was settled on after much deliberation). For those unfamiliar, a brief overview.

Over the past decade, important changes have begun to take root in the philanthropy/altruist sector:

- Organizations like Givewell, as well as a growing number of foundations like the Gates Foundation, are shifting the discussion of giving towards efficiency and evidence.
- Groups like Giving What We Can and Bolder Giving are encouraging people to incorporate philanthropy into their lifestyle. You can donate 10% or more of your income and still be among the richest people on the planet, living a satisfying life.
- The organization 80,000 Hours is promoting high impact career choice. You'll spent thousands of hours at your job. You can accomplish dramatically more good for the world if you optimize for it.

Above all, serious discussion is slowly mounting towards an incredibly important question - if you want to have the biggest impact you possibly can, what do you do?

Donating to provably efficient charities is an obvious first step, but more is possible. Systemic changes can have a powerful impact. New technologies have the potential to radically improve lives - as well as the capacity to destroy life as we know it. The Singularity Institute, the Future of Humanity Institute, Givewell and others are all in the process of grappling with this problem. I think it's fair to say that the Less Wrong community has had a noteworthy impact on the discussion.

We believe it's important that more people consider this question, and work on both the meta-tasks of comparing potential high impact causes, as well as the object-level tasks that follow.

A New Kind of Community

These ideas have been spreading. The seeds have been sown for a new kind of movement, which we believe has the potential to change the world on a scale rarely seen - at least not in a deliberate fashion. The Effective Altruism movement is growing slowly, but we think it's time for it to explode into something powerful and good.
In many ways this is not unlike the existing Less Wrong community. The NYC Less Wrong meetup has had a profound impact on me, personally. I've learned to explore important new ideas, think rigorously. I've learned the value of having likeminded people to share both important problems and my day to day experiences with. Most importantly, I've developed a sense of agency - I've realized I can personally cause big things to happen.
Less Wrong is about general rationality, which people can apply to numerous areas. There's tremendous value to having that, without attaching it to any cause or even meta-cause. But there's room for more than one community (truth be told I think everyone should have at least two tribes that don't fully intersect). There's an Eliezer quote I've been thinking about lately:

> "*Should the Earth last so long, I would like to see, as the form of rationalist communities, taskforces focused on all the work that needs doing to fix up the world.*"

Among the most valuable things the Less Wrong community has taught is the importance of... well, community. For Effective Altruism to be successful as a movement and a lifestyle, it needs people working together who share a passion for it, a commitment to intellectual rigor, and a sense of humor. People who can help each other grow, collaborate on important projects, and more.

THINK. The High Impact Network. Ready to launch this fall.

After just two months of work, we have approximately 30 volunteers and 6 directors, putting an average of 170 hours per week into THINK. Twenty meetups are gearing up to launch, with

discussions going to set up another thirty. Our English-speaking Facebook group has 103 members as I write this, and in just a week the Swedish-speaking group based in Stockholm went from 3 to 57 members.
This is just the beginning. We're ready to start tackling the world's biggest problems, and we hope you are too.

---

[1] Where by "help 'people'" we mean "and animals too." Depending on your ethical framework. Probably not including clams. Quite possibly including future sentient beings of various sorts. It's complicated. Come to a meetup, we'll talk about it.

Decision Theories, Part 3.5: Halt, Melt and Catch Fire

Followup to: Decision Theories: A Semi-Formal Analysis, Part III

UPDATE: As it turns out, rumors of Masquerade's demise seem to have been greatly exaggerated. See this post for details and proofs!

I had the chance, over the summer, to discuss the decision theory outlined in my April post with a bunch of relevantly awesome people. The sad part is, **there turned out to be a fatal flaw** once we tried to formalize it properly. I'm laying it out here, not with much hope that there's a fix, but because sometimes false starts can be productive for others.

Since it's not appropriate to call this decision theory TDT, I'm going to use a name suggested in one of these sessions and call it "Masquerade", which might be an intuition pump for how it operates. So let's first define some simple agents called "masks", and then define the "Masquerade" agent.

Say that our agent has actions a_1, ... , a_n, and the agent it's facing in this round has actions b_1, ... , b_m. Then for any triple (b_i, a_j, a_k), we can define a simple agent Mask_{ijk} which takes in its opponent's source code and outputs an action:

```
def Mask_ijk(opp_src):
 look for proof that Opp(Mask_ijk) = b_i
 if one is found, then output a_j
 otherwise, output a_k
```

(This is slightly less general than what I outlined in my post, but it'll do for our purposes. Note that there's no need for a_j and a_k to be distinct, so constant strategies fall under this umbrella as well.)

A key example of such an agent is what we might call FairBot: on a Prisoner's Dilemma, FairBot tries to prove that the other agent cooperates against FairBot, and if it finds such a proof, then it immediately cooperates. If FairBot fails to find such a proof, then it defects. (An important point is that if FairBot plays against itself and both have sufficiently strong deductive capacities, then a short proof of one's cooperation gives a slightly longer proof of the other's cooperation, and thus in the right circumstances we have mutual cooperation via Löb's Theorem.)

The agent Masquerade tries to do better than any individual mask (note that FairBot foolishly cooperates against CooperateBot when it could trivially do better by defecting). My original formulation can be qualitatively described as trying on different masks, seeing which one fares the best, and then running a "sanity check" to see if the other agent treats Masquerade the same way it treats that mask. The pseudocode looked like this:

```
def Masquerade(opp_src):
 for each (i,j,k), look for proofs of the form "Mask_ijk gets utility u against Opp"
 choose (i,j,k) corresponding to the largest such u found
 look for proof that Opp(Masquerade) = Opp(Mask_ijk)
```

```
if one is found, then output the same thing as Mask_ijk(Opp)
otherwise, output a default action
```

(The default should be something safe like a Nash equilibrium strategy, of course.)

Intuitively, when Masquerade plays the Prisoner's Dilemma against FairBot, Masquerade finds that the best utility against FairBot is achieved by some mask that cooperates, and then Masquerade's sanity-check is trying to prove that FairBot(Masquerade) = C as FairBot is trying to prove that Masquerade(FairBot) = C, and the whole Löbian circus goes round again. Furthermore, it's intuitive that when Masquerade plays against another Masquerade, the first one notices the proof of the above, and finds that the best utility against the other Masquerade is achieved by FairBot; thus both pass to the sanity-check stage trying to imitate FairBot, both seek to prove that the other cooperate against themselves, and both find the Löbian proof.

So what's wrong with this intuitive reasoning?

Problem: A deductive system can't count on its own consistency!

Let's re-examine the argument that Masquerade cooperates with FairBot. In order to set up the Löbian circle, FairBot needs to be able to prove that Masquerade selects a mask that cooperates with FairBot (like CooperateBot or FairBot). There are nice proofs that each of those masks attains the mutual-cooperation payoff against FairBot, but we also need to be sure that some other mask won't get the very highest (I defect, you cooperate) payoff against FairBot. Now you and I can see that this must be true, because FairBot simply can't be exploited that way. But crucially, *FairBot can't deduce its own inexploitability* without thereby becoming exploitable (for the same Gödelian reason that a formal system can't prove its own consistency unless it is actually inconsistent)!

Now, the caveats to this are important: if FairBot's deductive process is sufficiently stronger than the deductive process that's trying to exploit it (for example, FairBot might have an oracle that can answer questions about Masquerade's oracle, or FairBot might look for proofs up to length 2^N while Masquerade only looks up to length N), then it can prove (by exhaustion if nothing else) that Masquerade will select a cooperative mask after all. But since Masquerade needs to reason about Masquerade at this level, this approach goes nowhere. (At first, I thought that having a weaker oracle for Masquerade's search through masks, and a stronger oracle both for each mask and for Masquerade's sanity-check, would solve this. But that doesn't get off the ground: the agent thus defined attains mutual cooperation with FairBot, but not with itself, because the weaker oracle can't prove that it attains mutual cooperation with FairBot.)

Another caveat is the following: FairBot may not be able to rule out the provability of some statement we know is false, but (given a large enough deductive capacity) it can prove that a certain result is the first of its kind in a given ordering of proofs. So if our agents act immediately on the first proof they find, then we could make a version of Masquerade work... as long as each search *does* find a proof, and as long as *that* fact is provable by the same deduction system. But there's an issue with this: two masks paired against each other won't necessarily have provable outcomes!

Let's consider the following mask agent, which we'll call AntiFairBot: it searches for a proof that its opponent cooperates against it, and it *defects* if it finds one; if it doesn't find such a proof, then it *cooperates*. This may not be a very optimal agent, but it has one interesting property: if you pit AntiFairBot against FairBot, and the two of them use equivalent oracles, then it takes an oracle stronger than either to deduce what the two of them will do! Thus, Masquerade can't be sure that AntiFairBot won't get the highest payoff against FairBot (which of course it won't) unless it uses a stronger deduction system for the search through masks than FairBot uses for its proof search (which would mean that FairBot won't be able to tell what mask Masquerade picks).

I tried to fix this by iterating over only some of the masks; after all, there's no realistic opponent against whom AntiFairBot is superior to both FairBot and DefectBot. Unfortunately, at this point I realized two things: in order to play successfully against a reasonable range of opponents on the Prisoner's Dilemma, **Masquerade needs to be able to imitate at least both FairBot and DefectBot**; and **FairBot cannot prove that FairBot defects against DefectBot**. (There are variants of FairBot that *can* do so, e.g. it could search both for proofs of cooperation and proofs of defection and playing symmetrically if it finds one, but this variant is no longer guaranteed to cooperate against itself!)

If there are any problems with this reasoning, or an obvious fix that I've missed, please bring it to my attention; but otherwise, I've decided that my approach has failed drastically enough that it's time to do what Eliezer calls "halt, melt, and catch fire". The fact that Löbian cooperation works is enough to keep me optimistic about formalizing this side of decision theory in general, but the ideas I was using seem insufficient to succeed. (Some variant of "playing chicken with my deductive system" might be a crucial component.)

Many thanks to all of the excellent people who gave their time and attention to this idea, both on and offline, especially Eliezer, Vladimir Slepnev, Nisan, Paul Christiano, Critch, Alex Altair, Misha Barasz, and Vladimir Nesov. Special kudos to Vladimir Slepnev, whose gut intuition on the problem with this idea was immediate and correct.

An angle of attack on Open Problem #1

~~**There is a problem with the proof here and I have to think about whether I can fix it.** Thanks to vi21maobk9vp for pointing me in the right direction!~~ ***I have posted a new and hopefully correct proof attempt.*** *Thanks again to vi21maobk9vp!*

In his talk on open problems in Friendly AI, Eliezer's first question is how, given Löb's theorem, an AI can replace itself with a better expected utility maximizer that believes in as much mathematics as the original AI. I know exactly one trick for that sort of problem, so I decided to try that on a toy variant. To my surprise, it more or less just worked. Therefore:

Professor Quirrell proposes a game. You start with a score of one. Professor Quirrell moves first, by choosing a computer program and showing you its source code. You then have three options: Take your winnings; double down; or self-destruct.

If you take your winnings, the game ends, and your score is converted to Quirrell points.

If you self-destruct, the game ends, your score is lost, you'll be sent to bed without dinner, you'll lose 150 House points, Rita Skeeter will write a feature alleging that you're a Death Eater, and Professor Quirrell will publicly critique your performance. You are advised not to pick this option.

If you double down, your score doubles, and you advance to the next round. Professor Quirrell again moves first by choosing a computer program. Then, it's your turn—except that this time, you don't get to choose your move yourself: instead, it'll be chosen by Professor Quirrell's program from the previous round.

Professor Quirrell will endeavor to present an *educational* sequence of programs.

*

The idea is that Quirrell will make you self-destruct if he possibly can, so you must only accept programs that don't self-destruct, that accept only programs that don't self-destruct, that accept only programs that only accept—etc. That's supposed to capture one essential aspect of Eliezer's problem, namely how to make sure that a proposed rewrite doesn't destroy your values, while ignoring the complications due to a different aspect, namely comparing the expected values before and after the rewrite. In Quirrell's game, there are safe and unsafe rewrites, and you should always double down on a safe one and take your winnings when presented with an unsafe one. Let's look at some interesting programs that we could recognize as safe. [~~And to deal with the possibility of an infinite sequence of double-downs, let's stipulate a small but positive chance each round that Quirrell will end the game and pay you even if your program chose to double down.~~ **ETA:** Luke A. Somers points out that this provision isn't necessary.]

Let PA(0) := Peano Arithmetic, and PA(n+1) := PA(n) + for all formulas 'C': "if PA(n) proves 'C', then C". Define AI(n) by

```
def AI(n)(p):
    Look at the first 3^^^3 theorems of PA(n).
    if (one of them says "p is safe"):
        double down
    else:
        take winnings
```

If the current round is controlled by AI(7) and Quirrell hands it AI(6) as input, then AI(7) will double down: PA(7) will conclude that AI(6) will only double down on programs that PA(6) proves to be safe, which implies that they *are* safe. But if Quirrell *hands AI(7) to itself*, this reasoning doesn't go through, because PA(7) cannot use "PA(7) proves p is safe" to conclude "p is safe". So even if Quirrell is nice enough to choose AI(n)'s in a decreasing sequence, our windfall will end after a constant number rounds, because (***sinister drumroll***) AI(0) has run out of math to believe in, so to speak. That's precisely the problem Eliezer explains in his talk.

By quining, we could write an AI_Q which will recognize itself as safe, so if Quirrell chooses AI_Q over and over again, we'll be doubling down for longer and longer times. But that insight won't help us with writing interesting self-modifying AIs. [**ETA:** Wei Dai has managed to push the limits of using quining into interesting territory.] Is there something that can use non-specialized reasoning, like the AI(n), to recognize an indefinitely long sequence of variants of itself as safe?

*

Define "p is safe for n steps" (p a program) to mean that there is no m, $0 < m <= n$, and sequence of programs $p=p_0, p_1, ..., p_m$ such that (a) for $0 < i < m$, p_{i-1} will double down when fed p_i, and (b) p_{m-1} will ***self-destruct*** when fed p_m. Clearly this can be formalized in the language of Peano Arithmetic.

Now let's extend the language of PA by a constant symbol K, and define PA_K := Peano Arithmetic (actually, the natural extension of PA to this new language) + for all formulas 'C' of the base language: "if PA(K) proves 'C', then C". Finally, define AI_K by

```
def AI_K(p):
    Look at the first 3^^^3 theorems of PA_K.
    if (one of them says "p is safe for K steps"):
        double down
    else:
        take winnings
```

I claim that AI_K is safe, and furthermore, $AI_K(AI_K)$ will double down, and so will $AI_K(p)$ where p is some trivial modification of AI_K like changing 3^^^3 to 4^^^^4.

*

Fair warning: now comes the technical section of this post.

Proof sketch. Let inst(n,x) be the meta-level function that takes a sentence or proof x in the extended language and substitutes the numeral n (i.e., the unary number in the language of PA that encodes n) for all occurrences of the constant symbol K. Note that if x is a valid proof in PA_K, then inst(n,x) will be a valid proof in PA(n+1). Of course, what statement it is a proof *of* will depend on n. In particular, if x proves "p is safe for K steps", then inst(n,x) is a PA(n+1) proof of "p is safe for n steps". Since this is

argument works for all n, p will in fact be safe [and we can formalize this in PA_ω, if we want to]. This shows that AI_K is safe, as claimed.

Formalizing the above argument in Peano Arithmetic, and writing $inst_{PA}(n,x)$ for the object-level encoding of the meta-level function, we can prove: "For all extended-language sentences 'C', if PA_K proves 'C', then for all n, PA(n+1) proves $inst_{PA}(n,'C')$." By substituting K-1 for n, it follows that for all exteded-language sentences 'C', PA_K proves "If PA_K proves 'C', and K>0, then PA(K) proves $inst_{PA}(K\text{-}1,'C')$". Now, for 'D' := inst(K-1,'C'), PA can prove that "$inst_{PA}(K\text{-}1,'C')$ = 'D'"; thus, by virtue of its extra axioms, PA_K concludes: "If PA_K proves 'C', and K>0, then inst(K-1,C)".

What we must show is that PA_K proves "AI_K is safe for K steps" through some argument that doesn't actually involve simulating $AI_K(AI_K)$ (because in that case, we wouldn't expect to find the proof among the first 3^^^3 theorems). First, note that PA already ("quickly") proves "If $AI_K(p)$ doubles down, then PA_K proves 'p is safe for K steps'". In PA_K, it follows that "If $AI_K(p)$ doubles down, and K>0, then inst(K-1, 'p is safe for K steps')". But this is just a meta-level expression for the sentence "If $AI_K(p)$ doubles down, and K>0, then p is safe for K-1 steps"—and at that point, the reflective magic is over, and we only need to establish by "ordinary" PA reasoning that if AI_K will only accept programs that are safe for K-1 steps, then AI_K is safe for K steps.

Q.E.D.

Note that this argument does not depend on the number of theorems AI_K looks at, because it only depends on the fact that if $AI_K(p)$ *does* double down, then there is *some* PA_K proof of p's (K-1)-step safety.

*

The real question is whether this is just a hack or can tell us something about how to approach a solution to the real thing. It could very well be that this is one of the cheap tricks Eliezer and Marcello tried that don't solve the core problem (as Eliezer explains in his talk). Certainly the proof seems to be more tangled with the rounds structure of Quirrell's game than I find elegant. Also, I'm not at all sure that the key proof idea noted in the previous paragraph still works when we go from Quirrell's game to expected utility maximization.

However, as I said at the beginning of this post, I know exactly one trick for dealing with problems of this sort—by which I mean, trying to do something that seems impossible due to a diagonalization proof. It's well-appreciated that you can usually avoid diagonalization by passing from a single collection of things to a hierarchy of things (we can avoid Cantor by concluding that there are multiple infinite cardinalities; Gödel and Löb, by the PA(n) hierarchy; Turing, by a hierarchy of halting oracles; Tarski, by a hierarchy of truth predicates; and so on). It's less well appreciated, but I think true (though fuzzy), that many fields manage to circumvent the effects of diagonalization a bit further by considering objects that in some sense live on multiple levels of the hierarchy at the same time. I'd call that "the parametric polymorphism trick", perhaps.

In this post, we met PA_K, which can be interpreted as PA(n), for any n. I haven't tried it in detail, but something very similar should be possible for Tarski's truth predicates. A sufficiently powerful total programming language cannot have a self-interpreter, but

you should be able to have a constant symbol K, a single interpreter code file, and a hierarchy of semantics such that according to the K=n+1 semantics, the interpreter implements the K=n semantics. In Church's simple theory of types, you can't apply a function to itself, but in polymorphic variants, you can instantiate (id : α -> α) to (id : (α -> α) -> (α -> α)), so that id(id) makes sense. Set-theoretic models of the untyped lambda calculus need to deal with the fact that if a set S is isomorphic to the function space (S -> T), then S is either empty or has only one element; the usual solution would take me a bit more than one sentence to explain, but it's always struck me as related to what happens in the polymorphic type theories. Looking a bit farther afield, if you're a set theory platonist, if $\alpha<\beta<\gamma$ and if the von Neumann levels $\mathbf{V}_\alpha$, $\mathbf{V}_\beta$ and $\mathbf{V}_\gamma$ are all models of ZFC, then the ZFC proof that "if there is a set model of ZFC, then ZFC is consistent" can be interpreted in $\mathbf{V}_\beta$, where it applies to $\mathbf{V}_\alpha$, and it can also be interpreted in $\mathbf{V}_\gamma$, where it applies to both $\mathbf{V}_\alpha$ and $\mathbf{V}_\beta$. And so on. It may be that there's a good solution to the AI rewrite problem and it has nothing to do with this type of trick at all, but it seems at least worthwhile to look in that direction.

[Actually, come to think of it, I *do* know a second trick for circumventing diagonalization, exemplified by passing from total to partial recursive functions and from two- to three-valued logics, but in logic, that one usually makes the resulting systems too weak to be interesting.]

*

Three more points in closing. First, sorry for the informality of the proof sketch! It would be very much appreciated if people would go through it and point out unclear things/problems/corrections. Also, I'm hoping to make a post at some point that gives a more intuitive explanation for *why* this works.

[**ETA:** vi21maobk9vp points out that the following note may lead to unhelpful confusion; perhaps best skip this or read our discussion in the thread above that comment.] Second, provability in PA_K in some sense says that a sentence is provable in PA(n), for all n. In particular, PA_K is conservative over PA(1), since if PA_K proves a sentence C in the base language, then PA(1) proves inst(1,C), which is just C; in some sense, this makes PA_K rather weak. If we don't want this, we could make a variant where provability implies says there is some m such that the sentence is provable in PA(n) for all n>m. To do this, we'd use a trick usually employed to show that there are non-standard models of the natural numbers: Add to PA_K the axioms "K>1", "K>2", "K>3", and so on. This is consistent by the Compactness Theorem, because any concrete proof can only use a finite number of these axioms. But in this extended system, we can prove anything that we can prove in any PA(n).

Third, I chose the particular definition of PA_K because it made my particular proof simpler to write. Looking only at the definitions, I would find it more natural to make PA_K conservative over PA(0) by using "if K>0 and PA(K-1) proves 'C', then C".

Natural Laws Are Descriptions, not Rules

Laws as Rules

We speak casually of the laws of nature *determining* the distribution of matter and energy, or *governing* the behavior of physical objects. Implicit in this rhetoric is a metaphysical picture: the laws are *rules* that constrain the temporal evolution of stuff in the universe. In some important sense, the laws are prior to the distribution of stuff. The physicist Paul Davies expresses this idea with a bit more flair: "[W]e have this image of really existing laws of physics ensconced in a transcendent aerie, lording it over lowly matter." The origins of this conception can be traced back to the beginnings of the scientific revolution, when Descartes and Newton established the discovery of laws as the central aim of physical inquiry. In a scientific culture immersed in theism, it was unproblematic, even natural, to think of physical laws as rules. They are rules laid down by God that drive the development of the universe in accord with His divine plan.

Does this prescriptive conception of law make sense in a secular context? Perhaps if we replace the divine creator of traditional religion with a more naturalist-friendly lawgiver, such as an ur-simulator. But what if there is no intentional agent at the root of it all? Ordinarily, when I think of a physical system as constrained by some rule, it is not the rule itself doing the constraining. The rule is just a piece of language; it is an *expression* of a constraint that is actually enforced by interaction with some other physical system -- a programmer, say, or a physical barrier, or a police force. In the sort of picture Davies presents, however, it is the rules themselves that enforce the constraint. The laws lord it over lowly matter. So on this view, the fact that all electrons repel one another is explained by the existence of some external entity, not an ordinary physical entity but a law of nature, that somehow *forces* electrons to repel one another, and this isn't just short-hand for God or the simulator forcing the behavior.

I put it to you that this account of natural law is utterly mysterious and borders on the nonsensical. How exactly are abstract, non-physical objects -- laws of nature, living in their "transcendent aerie" -- supposed to interact with physical stuff? What is the mechanism by which the constraint is applied? Could the laws of nature have been different, so that they forced electrons to attract one another? The view should also be anathema to any self-respecting empiricist, since the laws appear to be idle danglers in the metaphysical theory. What is the difference between a universe where all electrons, as a matter of contingent fact, attract one another, and a universe where they attract one another because they are compelled to do so by the really existing laws of physics? Is there any test that could distinguish between these states of affairs?

Laws as Descriptions

There are those who take the incoherence of the secular prescriptive conception of laws as reason to reject the whole concept of laws of nature as an anachronistic holdover from a benighted theistic age. I don't think the situation is that dire.

Discovering laws of nature is a hugely important activity in physics. It turns out that the behavior of large classes of objects can be given a unified compact mathematical description, and this is crucial to our ability to exercise predictive control over our environment. The significant word in the last sentence is "description". A much more congenial alternative to the prescriptive view is available. Instead of thinking of laws as rules that have an existence above and beyond the objects they govern, think of them as particularly concise and powerful descriptions of regular behavior.

On this descriptive conception of laws, the laws do not exist independently in some transcendent realm. They are not prior to the distribution of matter and energy. The laws are just descriptions of salient patterns in that distribution. Of course, if this is correct, then our talk of the laws governing matter must be understood as metaphorical, but this is a small price to pay for a view that actually makes sense. There may be a concern that we are losing some important explanatory ground here. After all, on the prescriptive view the laws of nature *explain* why all electrons attract one another, whereas on the descriptive view the laws just restate the fact that all electrons attract one another. But consider the following dialogue:

A: Why are these two metal blocks repelling each other?

B: Because they're both negatively charged, which means they have an excess of electrons, and electrons repel one another.

A: But why do electrons repel one another?

B: Because like charges always repel.

A: But why is that?

B: Because if you do the path integral for the electromagnetic field (using Maxwell's Lagrangian) with source terms corresponding to two spatially separated lumps of identical charge density, you will find that the potential energy of the field is greater the smaller the spatial separation between the lumps, and we know the force points in the opposite direction to the gradient of the potential energy.

A: But why are the dynamics of the electromagnetic field derived from Maxwell's Lagrangian rather than some other equation? And why does the path integral method work at all?

B: BECAUSE IT IS THE LAW.

Is the last link in this chain doing any explanatory work at all? Does it give us any further traction on the problem? *B* might as well have ended that conversation by saying "Well, that's just the way things are." Now, laws of nature do have a privileged role in physical explanation, but that privilege is due to their simplicity and generality, not to some mysterious quasi-causal power they exert over matter. The fact that a certain generalization is a law of nature does not *account* for the truth and explanatory power of the generalization, any more than the fact that a soldier has won the Medal of Honor accounts for his or her courage in combat. Lawhood is a *recognition* of the generalization's truth and explanatory power. It is an honorific; it doesn't confer any further explanatory oomph.

The Best System Account of Laws

David Lewis offers us a somewhat worked out version of the descriptive conception of law. Consider the set of all truths about the world expressible in a particular language. We can construct deductive systems out of this set of propositions by picking out some of the propositions as axioms. The logical consequences of these axioms are the theorems of the deductive system. These deductive systems compete with one another along (at least) two dimensions: the *simplicity* of the axioms, and the *strength* or *information content* of the system as a whole. We prefer systems that give us more information about the world, but this greater strength often comes at the cost of simplicity. For instance, a system whose axioms comprised the entire set of truths about the world would be maximally strong, but not simple at all. Conversely, a system whose only axiom is something like "Stuff happens" would be pretty simple, but very uninformative. What we are looking for is the appropriate balance of simplicity and strength [1].

According to Lewis, the laws of nature correspond to the axioms of the deductive system that best balances simplicity and strength. He does not provide a precise algorithm for evaluating this balance, and I don't think his proposal should be read as an attempt at a technically precise decision procedure for lawhood anyway. It is more like a heuristic picture of what we are doing when we look for laws. We are looking for simple generalizations that can be used to deduce a large amount of information about the world. Laws are highly compressed descriptions of broad classes of phenomena. This view evidently differs quite substantially from the Davies picture I presented at the beginning of this post. On Lewis's view, the collection of particular facts about the world determines the laws of nature, since the laws are merely compact descriptions of those facts. On Davies's view, the determination runs the other way. The laws are independent entities that determine the particular facts about the world. Stuff in the world is arranged the way it is because the laws compelled that arrangement.

One last point about Lewis's account. Lewis acknowledges that there is an important language dependence in his view of laws. If we ignore this, we get absurd results. For instance, consider a system whose only axiom is "For all x, x is F" where "F" is defined to be a predicate that applies to all and only events that occur in this world. This axiom is maximally informative, since it rules out all other possible worlds, and it seems exceedingly simple. Yet we wouldn't want to declare it a law of nature. The problem, obviously, is that all the complexity of the axiom is hidden by our choice of language, with this weird specially rigged predicate. To rule out this possibility, Lewis specifies that all candidate deductive systems must employ the vocabulary of fundamental physics.

But we could also regard lawhood as a 2-place function which maps a proposition and vocabulary pair to "True" if the proposition is an axiom of the best system in that vocabulary and "False" otherwise. Lewis has chosen to curry this function by fixing the vocabulary variable. Leaving the function uncurried, however, highlights that we could have different laws for different vocabularies and, consequently, for different levels of description. If I were an economist, I wouldn't be interested (at least not *qua* economist) in deductive systems that talked about quarks and leptons. I would be interested in deductive systems that talked about prices and demand. The best system for this coarser-grained vocabulary will give us the laws of economics, distinct from the laws of physics.

Lawhood Is in the Map, not in the Territory

There is another significant difference between the descriptive and prescriptive accounts that I have not yet discussed. On the Davies-style conception of laws as rules, lawhood is an element of reality. A law is a distinctive beast, an abstract entity perched in a transcendent aerie. On the descriptive account, by comparison, lawhood is part of our map, not the territory. Note that I am not saying that the *laws themselves* are a feature of the map and not the territory. Laws are just particularly salient redundancies, ones that permit us to construct useful compressed descriptions of reality. These redundancies are, of course, out there in the territory. However, the fact that certain regularities are especially useful for the organization of knowledge is at least partially dependent on facts about us, since we are the ones doing the organizing in pursuit of our particular practical projects. Nature does not flag these regularities as laws, we do.

This realization has consequences for how we evaluate certain forms of reductionism. I should begin by noting that there is a type of reductionism I tentatively endorse and that I think is untouched by these speculations. I call this *mereological reductionism* [2]; it is the claim that all the stuff in the universe is entirely built out of the kinds of things described by fundamental physics. The vague statement is intentional, since fundamental physicists aren't yet sure what kinds of things they are describing, but the motivating idea behind the view is to rule out the existence of immaterial souls and the like. However, reductionists typically embrace a stronger form of reductionism that one might label *nomic reductionism* [3]. The view is that the fundamental laws of physics are the only *really existant* laws, and that laws in the non-fundamental disciplines are merely convenient short-cuts that we must employ due to our computational limitations.

One appealing argument for this form of reductionism is the apparent superfluity of non-fundamental laws. Macroscopic systems are entirely built out of parts whose behavior is determined by the laws of physics. It follows that the behavior of these systems is also fixed by those fundamental laws. Additional non-fundamental laws are otiose; there is nothing left for them to do. Barry Loewer puts it like this: "Why would God make [non-fundamental laws] the day after he made physics when the world would go on exactly as if they were there without them?" If these laws play no explanatory role, Ockham's razor demands that we strike them from our ontological catalog, leaving only the fundamental laws.

I trust it is apparent that this argument relies on the prescriptive conception of laws. It assumes that real laws of nature *do stuff*; they push and pull matter and energy around. It is this implicit assumption that raises the overdetermination concern. On this assumption, if the fundamental laws of physics are already lording it over all matter, then there is no room for another locus of authority. However, the argument (and much of the appeal of the associated reductionist viewpoint) fizzles, if we regard laws as descriptive. Employing a Lewisian account, all we have are different best systems, geared towards vocabularies at different resolutions, that highlight different regularities as the basis for a compressed description of a system. There is nothing problematic with having different ways to compress information about a system. Specifically, we are not compelled by worries about overdetermination to declare one of these methods of compression to be *more real* than another. In response to Loewer's theological question, the proponent of the descriptive conception could say that God does not get to separately specify the non-fundamental and fundamental laws. By creating the pattern of events in space-time she implicitly fixes them all.

Nomic reductionism would have us believe that the lawhood of the laws of physics is part of the territory, while the lawhood of the laws of psychology is just part of our

map. Once we embrace the descriptive conception of laws, however, there is no longer this sharp ontological divide between the fundamental and non-fundamental laws. One reason for privileging the laws of physics is revealed to be the product of a confused metaphysical picture. However, one might think there are still other good reasons for privileging these laws that entail a reductionism more robust than the mereological variety. For instance, even if we accept that laws of physics don't possess a different ontological status, we can still believe that they have a prized position in the explanatory hierarchy. This leads to *explanatory reductionism*, the view that explanations couched in the vocabulary of fundamental physics are always better because fundamental physics provides us with more accurate models than the non-fundamental sciences. Also, even if one denies that the laws of physics themselves are pushing matter around, one can still believe that all the actual pushing and pulling there is, all the causal action, is described by the laws of physics, and that the non-fundamental laws do not describe genuine causal relations. We could call this kind of view *causal reductionism*.

Unfortunately for the reductionist, explanatory and causal reductionism don't fare much better than nomic reductionism. Stay tuned for the reasons why!

---

[1] Lewis actually adds a third desideratum, *fit*, that allows for the evaluation of systems with probabilistic axioms, but I leave this out for simplicity of exposition. I have tweaked Lewis's presentation in a couple of other ways as well. For his own initial presentation of the view, see *Counterfactuals*, pp. 72-77. For a more up-to-date presentation, dealing especially with issues involving probabilistic laws, see this paper (PDF).

[2] From the Greek *meros*, meaning "part".

[3] From the Greek *nomos*, meaning "law".