Best of LessWrong: November 2017

1. Hero Licensing
2. Sunset at Noon
3. Moloch's Toolbox (1/2)
4. Modesty and diversity: a concrete suggestion
5. Moloch's Toolbox (2/2)
6. The Copernican Revolution from the Inside
7. Announcing the AI Alignment Prize
8. The Archipelago Model of Community Standards
9. Mosquito killing begins
10. Security Mindset and Ordinary Paranoia
11. Cooperative model knob-turning
12. Preferences over non-rewards
13. Competitive Truth-Seeking
14. 'X is not about Y' is not about psychology
15. Open thread, November 13 - November 20, 2017
16. Clarify Your No's
17. The Happy Dance Problem
18. Living in an Inadequate World
19. Mapping Another's Universe
20. Security Mindset and the Logistic Success Curve
21. Reason as Attentional Prosthesis
22. Open thread, November 21 - November 28, 2017
23. Status Regulation and Anxious Underconfidence
24. Normative assumptions: answers, emotions, and narratives
25. Productivity: Working towards a summary of what we know
26. Zeroing Out
27. Rationalising humans: another mugging, but not Pascal's
28. Gears Level & Policy Level
29. Civility Is Never Neutral
30. Against Modest Epistemology
31. Qualitative differences
32. The Darwin Game
33. Blind Empiricism
34. The Right to be Wrong
35. Military AI as a Convergent Goal of Self-Improving AI
36. The Darwin Results
37. Stable agent, subagent-unstable
38. Big Advance in Infinite Ethics
39. Against Shooting Yourself in the Foot
40. Lizard Jockeying for Fun and Profit
41. XOR Blackmail & Causality
42. Remove Intercom?
43. The Darwin Pregame
44. USA v Progressive 1979 excerpt
45. Creating Welfare Biology: A Research Proposal
46. Our values are underdefined, changeable, and manipulable
47. The Journal of High Standards
48. Anthropic reasoning isn't magic
49. Confessions of a Slacker
50. The Mad Scientist Decision Problem

Best of LessWrong: November 2017

1. Hero Licensing
2. Sunset at Noon
3. Moloch's Toolbox (1/2)
4. Modesty and diversity: a concrete suggestion
5. Moloch's Toolbox (2/2)
6. The Copernican Revolution from the Inside
7. Announcing the AI Alignment Prize
8. The Archipelago Model of Community Standards
9. Mosquito killing begins
10. Security Mindset and Ordinary Paranoia
11. Cooperative model knob-turning
12. Preferences over non-rewards
13. Competitive Truth-Seeking
14. 'X is not about Y' is not about psychology
15. Open thread, November 13 - November 20, 2017
16. Clarify Your No's
17. The Happy Dance Problem
18. Living in an Inadequate World
19. Mapping Another's Universe
20. Security Mindset and the Logistic Success Curve
21. Reason as Attentional Prosthesis
22. Open thread, November 21 - November 28, 2017
23. Status Regulation and Anxious Underconfidence
24. Normative assumptions: answers, emotions, and narratives
25. Productivity: Working towards a summary of what we know
26. Zeroing Out
27. Rationalising humans: another mugging, but not Pascal's
28. Gears Level & Policy Level
29. Civility Is Never Neutral
30. Against Modest Epistemology
31. Qualitative differences
32. The Darwin Game
33. Blind Empiricism
34. The Right to be Wrong
35. Military AI as a Convergent Goal of Self-Improving AI
36. The Darwin Results
37. Stable agent, subagent-unstable
38. Big Advance in Infinite Ethics
39. Against Shooting Yourself in the Foot
40. Lizard Jockeying for Fun and Profit
41. XOR Blackmail & Causality
42. Remove Intercom?
43. The Darwin Pregame
44. USA v Progressive 1979 excerpt
45. Creating Welfare Biology: A Research Proposal
46. Our values are underdefined, changeable, and manipulable
47. The Journal of High Standards
48. Anthropic reasoning isn't magic
49. Confessions of a Slacker

50. The Mad Scientist Decision Problem

Hero Licensing

I expect most readers to know me either as MIRI's co-founder and the originator of a number of the early research problems in AI alignment, or as the author of *Harry Potter and the Methods of Rationality*, a popular work of *Harry Potter* fanfiction. I've described how I apply concepts in *Inadequate Equilibria* to various decisions in my personal life, and some readers may be wondering how I see these tying in to my AI work and my fiction-writing. And I do think these serve as useful case studies in inadequacy, exploitability, and modesty.

As a supplement to *Inadequate Equilibria*, then, the following is a dialogue that never took place—largely written in 2014, and revised and posted online in 2017.

---

i. Outperforming and the outside view

(*The year is 2010.* ELIEZER-2010 *is sitting in a nonexistent park in Redwood City, California, working on his laptop. A* PERSON *walks up to him.*)

PERSON: Pardon me, but are you Eliezer Yudkowsky?

ELIEZER-2010: I have that dubious honor.

PERSON: My name is Pat; Pat Modesto. We haven't met, but I know you from your writing online. What are you doing with your life these days?

ELIEZER-2010: I'm trying to write a nonfiction book on rationality. The blog posts I wrote on *Overcoming Bias*—I mean *Less Wrong*—aren't very compact or edited, and while they had some impact, it seems like a book on rationality could reach a wider audience and have a greater impact.

PAT: Sounds like an interesting project! Do you mind if I peek in on your screen and—

ELIEZER: (*shielding the screen*) —Yes, I mind.

PAT: Sorry. Um... I did catch a glimpse and that didn't look like a nonfiction book on rationality to me.

ELIEZER: Yes, well, work on that book was going very slowly, so I decided to try to write something else in my off hours, just to see if my general writing speed was slowing down to molasses or if it was this particular book that was the problem.

PAT: It looked, in fact, like *Harry Potter* fanfiction. Like, I'm pretty sure I saw the words "Harry" and "Hermione" in configurations not originally written by J. K. Rowling.

ELIEZER: Yes, and I currently seem to be writing it very quickly. And it doesn't seem to use up mental energy the way my regular writing does, either.

(*A* MYSTERIOUS MASKED STRANGER*, watching this exchange, sighs wistfully.*)

ELIEZER: Now I've just got to figure out why my main book-writing project is going so much slower and taking vastly more energy... There are *so many* books I could write, if I could just write everything as fast as I'm writing this...

PAT: Excuse me if this is a silly question. I don't mean to say that *Harry Potter* fanfiction is bad—in fact I've read quite a bit of it myself—but as I understand it, according to your basic philosophy the world is currently on fire and needs to be put out. Now given that this is true, why are you writing *Harry Potter* fanfiction, rather than doing something else?

ELIEZER: I am doing something else. I'm writing a nonfiction rationality book. This is just in my off hours.

PAT: Okay, but I'm asking why you are doing *this* particular thing in your off hours.

ELIEZER: Because my life is limited by mental energy far more than by time. I can currently produce this work very cheaply, so I'm producing more of it.

PAT: What I'm trying to ask is why, *even given* that you can write *Harry Potter* fanfiction very cheaply, you are writing *Harry Potter* fanfiction. Unless it really is true that the only reason is that you need to observe yourself writing quickly in order to understand the way of quick writing, in which case I'd ask what probability you assign to learning that successfully. I'm skeptical that this is *really* the best way of using your off hours.

ELIEZER: I'm skeptical that you have correctly understood the concept of "off hours." There's a reason they exist, and the reason isn't just that humans are lazy. I admit that Anna Salamon and Luke Muehlhauser don't require off hours, but I don't think they are, technically speaking, "humans."

(*The Mysterious Masked Stranger speaks for the first time.*)

STRANGER: Excuse me.

ELIEZER: Who are you?

STRANGER: No one of consequence.

PAT: And why are you wearing a mask?

STRANGER: Well, I'm definitely not a version of Eliezer from 2014 who's secretly visiting the past, if that's what you're thinking.

PAT: It's fair to say that's not what I'm thinking.

STRANGER: Pat and Eliezer-2010, I think the two of you are having some trouble communicating. The two of you actually disagree much more than you think.

PAT & ELIEZER: Go on.

STRANGER: If you ask Eliezer of February 2010 why he's writing *Harry Potter and the Methods of Rationality*, he will, indeed, respond in terms of how he expects writing *Methods* to positively impact his attempt to write *The Art of Rationality*, his attempt at a nonfiction how-to book. This is because we have—I mean, Eliezer has—a heuristic of planning on the mainline, which means that his primary justification for anything will be phrased in terms of how it positively contributes to a "normal" future timeline, not low-probability side-scenarios.

ELIEZER: Sure.

PAT: Wait, isn't your whole life—

ELIEZER: No.

STRANGER: Eliezer-2010 *also* has a heuristic that might be described as "never try to do anything unless you have a chance of advancing the Pareto frontier of the category." In other words, if he's expecting that some other work will be strictly better than his along all dimensions, it won't occur to Eliezer-2010 that this is something he should spend time on. Eliezer-2010 thinks he has the potential to do things that advance Pareto frontiers, so why would he consider a project that wasn't trying? So, off-hours or not, Eliezer wouldn't be working on this story if he thought it would be strictly dominated along every dimension by any other work of fanfiction, or indeed, any other book.

PAT: Um—

ELIEZER: I wouldn't put it in exactly those terms.

STRANGER: Yes, because when you say things like that out loud, people start saying the word "arrogance" a lot, and you don't fully understand the reasons. So you'll cleverly dance around the words and try to avoid that branch of possible conversation.

PAT: Is that true?

ELIEZER: It sounds to me like the Masked Stranger is trying to use the Barnum effect—like, most people would acknowledge that as a secret description of themselves if you asked them.

PAT: ..... I really, really don't think so.

ELIEZER: I'd be surprised if it were less than 10% of the population, seriously.

STRANGER: Eliezer, you'll have a somewhat better understanding of human status emotions in 4 years. Though you'll still only go there when you have a point to make that can't be made any other way, which in turn will be unfortunately often as modest epistemology norms propagate through your community. But anyway, Pat, the fact that Eliezer-2010 has spent any significant amount of time on *Harry Potter and the Methods of Rationality* indeed lets you infer that Eliezer-2010 thinks *Methods* has a chance of being *outstanding* along some key dimension that interests him—of

advancing the frontiers of what has ever been done—although he might hesitate to tell you that before he's actually done it.

ELIEZER: Okay, yes, that's true. I'm unhappy with the treatment of supposedly "intelligent" and/or "rational" characters in fiction and I want to see it done right just once, even if I have to write the story myself. I have an explicit thesis about what's being done wrong and how to do it better, and if this were not the case then the prospect of writing *Methods* would not interest me as much.

STRANGER: (*aside*) There's so much civilizational inadequacy in our worldview that we hardly even notice when we invoke it. Not that this is an alarming sign, since, as it happens, we do live in an inadequate civilization.

ELIEZER: (*continuing to Pat*) However, the reason I hold back from saying in advance what *Methods* might accomplish isn't just modesty. I'm genuinely unsure that I can make *Methods* be what I think it can be. I don't want to promise more than I can deliver. And since one should first plan along the mainline, if investigating the conditions under which I can write quickly weren't a sufficiently important reason, I wouldn't be doing this.

STRANGER: (*aside*) I have some doubts about that alleged justification in retrospect, though it wasn't stupid.

PAT: Can you say more about how you think your *Harry Potter* story will have outstandingly "intelligent" characters?

ELIEZER: I'd rather not? As a matter of literature, I should show, not tell, my thesis. Obviously it's not that I think that my characters are going to learn fifty-seven languages because they're super-smart. I think most attempts to create "intelligent characters" focus on surface qualities, like how many languages someone has learned, or they focus on stereotypical surface features the author has seen in other "genius" characters, like a feeling of alienation. If it's a movie, the character talks with a British accent. It doesn't seem like most such authors are aware of Vinge's reasoning for why it should be *hard* to write a character that is smarter than the author. Like, if you know exactly where an excellent chessplayer would move on a chessboard, you must be at least that good at playing chess yourself, because you could always just make that move. For exactly the same reason, it's hard to write a character that's more rational than the author.

I don't think the concept of "intelligence" or "rationality" that's being used in typical literature has anything to do with discerning good choices or making good predictions. I don't think there *is* a standard literary concept for characters who excel at cognitive optimization, distinct from characters who just win because they have a magic sword in their brains. And I don't think most authors of "genius" characters respect their supposed geniuses enough to really put themselves in their shoes—to really feel what their inner lives would be like, and think beyond the first cliche that comes to mind. The author still sets themselves above the "genius," gives the genius some kind of obvious stupidity that lets the author maintain emotional distance...

STRANGER: (*aside*) Most writers have a hard time conceptualizing a character who's genuinely smarter than the author; most futurists have a hard time conceptualizing genuinely smarter-than-human AI; and indeed, people often neglect the hypothesis that particularly smart human beings will have already taken into account all the factors that they consider obvious. But with respect to sufficiently competent *individuals* making decisions that they can make on their own cognizance—as

opposed to any larger bureaucracy or committee, or the collective behavior of a field—it *is* often appropriate to ask if they might be smarter than you think, or have better justifications than are obvious to you.

PAT: Okay, but supposing you can write a book with intelligent characters, how does *that* help save the world, exactly?

ELIEZER: Why are you focusing on the word "intelligence" instead of "rationality"? But to answer your question, nonfiction writing conveys facts; fiction writing conveys *experiences.* I'm worried that my previous two years of nonfiction blogging haven't produced nearly enough transfer of real cognitive skills. The hope is that writing about the inner experience of someone trying to be rational will convey things that I can't easily convey with nonfiction blog posts.

STRANGER: (*laughs*)

ELIEZER: What is it, Masked Stranger?

STRANGER: Just... you're so very *modest.*

ELIEZER: You're saying this to *me?*

STRANGER: It's sort of obvious from where I live now. So very careful not to say what you *really* hope *Harry Potter and the Methods of Rationality* will do, because you know people like Pat won't believe it and can't be persuaded to believe it.

PAT: This guy is weird.

ELIEZER: (*shrugging*) A lot of people are.

PAT: Let's ignore him. So you're presently investing a lot of hours—

ELIEZER: But surprisingly little mental energy.

STRANGER: Where I come from, we would say that you're investing surprisingly few spoons.

PAT: —but still a lot of hours, into crafting a *Harry Potter* story with, you hope, exceptionally rational characters. Which will cause some of your readers to absorb the experience of being rational. Which you think eventually ends up important to saving the world.

ELIEZER: Mm, more or less.

PAT: What do you think the outside view would say about—

ELIEZER: Actually, I think I'm about out of time for today. (*Starts to close his laptop.*)

STRANGER: Wait. Please stick around. Can you take my word that it's important?

ELIEZER: ...all right. I suppose I don't have very much experience with listening to Masked Strangers, so I'll try that and see what happens.

PAT: What did I say wrong?

STRANGER: You said that the conversation would never go anywhere helpful.

ELIEZER: I wouldn't go that far. It's true that in my experience, though, people who use the phrase "outside view" usually don't offer advice that I think is true, and the conversations take up a lot of mental energy—spoons, you called them? But since I'm taking the Masked Stranger's word on things and trying to continue, fine. What do you think the outside view has to say about the *Methods of Rationality* project?

PAT: Well, I was just going to ask you to consider what the average story with a rational character in it accomplishes in the way of skill transfer to readers.

ELIEZER: I'm not trying to write an average story. The whole point is that I think the average story with a "rational" character is screwed up.

PAT: So you think that your characters will be *truly* rational. But maybe those authors also think *their* characters are rational—

ELIEZER: (*in a whisper to the Masked Stranger*) Can I exit this conversation?

STRANGER: No. Seriously, it's important.

ELIEZER: Fine. Pat, your presumption is wrong. These hypothetical authors making a huge effort to craft rational characters don't actually exist. They don't realize that it *should* take an effort to craft rational characters; they're just regurgitating cliches about Straw Vulcans with very little self-perceived mental effort.

STRANGER: Or as I would phrase it: This is not one of the places where our civilization puts in enough effort that we should expect adequacy.

PAT: Look, I don't dispute that you can probably write characters more rational than those of the average author; I just think it's important to remember, on each occasion, that being wrong feels just like being right.

STRANGER: Eliezer, please tell him what you actually think of that remark.

ELIEZER: You do *not* remember on each occasion that "being wrong feels just like being right." You remember it on highly selective occasions where you are motivated to be skeptical of someone else. This feels just like remembering it on every relevant occasion, since, after all, every time you felt like you ought to think of it, you did. You just used a fully general counterargument, and the problem with arguments like that is that they provide no Bayesian discrimination between occasions where we are wrong and occasions where we are right. Like "but I have faith," "being wrong feels just like being right" is as easy to say on occasions when someone is right as on occasions when they are wrong.

STRANGER: There *is* a stage of cognitive practice where people should meditate on how the map is not the territory, especially if it's never before occurred to them that what feels like the universe of their immersion is actually their brain's reconstructed map of the true universe. It's just that Eliezer went through that phase while reading S. I. Hayakawa's *Language in Thought and Action* at age eleven or so. Once that lesson is fully absorbed internally, invoking the map-territory distinction as a push against ideas you don't like is (fully general) motivated skepticism.

PAT: Leaving that aside, there's this research showing that there's a very useful technique called "reference class forecasting"—

ELIEZER: I am aware of this.

PAT: And I'm wondering what reference class forecasting would say about your attempt to do good in the world via writing *Harry Potter* fanfiction.

ELIEZER: (*to the Masked Stranger*) *Please* can I run away?

STRANGER: No.

ELIEZER: (*sighing*) Okay, to take the question seriously as more than generic skepticism: If I think of the books which I regard as having well-done rational characters, their track record isn't bad. A. E. van Vogt's *The World of Null-A* was an inspiration to me as a kid. *Null-A* didn't just teach me the phrase "the map is not the territory"; it was where I got the idea that people employing rationality techniques ought to be awesome and if they weren't awesome that meant they were doing something wrong. There are a heck of a lot of scientists and engineers out there who were inspired by reading one of Robert A. Heinlein's hymns in praise of science and engineering—yes, I know Heinlein had problems, but the fact remains.

STRANGER: I wonder what smart kids who grew up reading *Harry Potter and the Methods of Rationality* as twelve-year-olds will be like as adults...

PAT: But surely van Vogt's *Null-A* books are an *exceptional* case of books with rationalist characters. My first question is, what reason do you have to believe *you* can do that? And my second question is, even given that you write a rational character as inspiring as a character in a Heinlein novel, how much impact do you think one character like that has on an average reader, and how many people do you think will read your *Harry Potter* fanfiction in the best case?

ELIEZER: To be honest, it feels to me like you're asking the wrong questions. Like, it would never occur to me to ask any of the questions you're asking now, in the course of setting out to write *Methods*.

STRANGER: (*aside*) That's true, by the way. None of these questions ever crossed my mind in the original timeline. I'm only asking them now because I'm writing the character of Pat Modesto. A voice like Pat Modesto is *not a productive voice to have inside your head,* in my opinion, so I don't spontaneously wonder what he would say.

ELIEZER: To produce the best novel I can, it makes sense for me to ask what other authors were doing wrong with their rational characters, and what A. E. van Vogt was doing right. I *don't* see how it makes sense for me to be nervous about whether I can do better than A. E. van Vogt, who had no better source to work with than Alfred Korzybski, decades before Daniel Kahneman was born. I mean, to be honest about what I'm really thinking: So far as I'm concerned, I'm already walking outside whatever so-called reference class you're inevitably going to put me in—

PAT: What?! What the heck does it mean to "walk outside" a reference class?

ELIEZER: —which doesn't guarantee that I'll succeed, because being outside of a reference class isn't the same as being better than it. It means that I don't draw conclusions from the reference class to myself. It means that I try, and see what happens.

PAT: You think you're just automatically better than every other author who's ever tried to write rational characters?

ELIEZER: No! Look, thinking things like that is just not how the inside of my head is organized. There's just the book I have in my head and the question of whether I can translate that image into reality. *My mental world is about the book, not about me.*

PAT: But if the book you have in your head implies that you can do things at a very high percentile level, relative to the average fiction author, then it seems reasonable for me to ask why you *already* think you occupy that percentile.

STRANGER: Let me try and push things a bit further. Eliezer-2010, suppose I told you that as of the start of 2014, *Methods* succeeded to the following level. First, it has roughly half a million words, but you're not finished writing it—

ELIEZER: Damn. That's disappointing. I must have slowed down a lot, and definitely haven't mastered the secret of whatever speed-writing I'm doing right now. I wonder what went wrong? Actually, why am I hypothetically continuing to write this book instead of giving up?

STRANGER: Because it's the most reviewed work of *Harry Potter* fanfiction out of more than 500,000 stories on fanfiction.net, has organized fandoms in many universities and colleges, has received at least 15,000,000 page views on what is no longer the main referenced site, has been turned by fans into an audiobook via an organized project into which you yourself put zero effort, has been translated by fans into many languages, is famous among the Caltech/MIT crowd, has its own daily-trafficked subreddit with 6,000 subscribers, is often cited as the most famous or the most popular work of *Harry Potter* fanfiction, is considered by a noticeable fraction of its readers to be literally the best book they have ever read, and on at least one occasion inspired an International Mathematical Olympiad gold medalist to join the alliance and come to multiple math workshops at MIRI.

ELIEZER: I like this scenario. It is weird, and I like weird. I would derive endless pleasure from inflicting this state of affairs on reality and forcing people to come to terms with it.

STRANGER: Anyway, what probability would you assign to things going at least that well?

ELIEZER: Hm... let me think. Obviously this *exact* scenario is improbable, because conjunctive. But if we partition outcomes according to whether they rank at least this high or better in my utility function, and ask how much probability mass I put into outcomes like that, then I think it's around 10%. That is, a success like this would come in at around the 90th percentile of my hopes.

PAT: (*incoherent noises*)

ELIEZER: Oh. Oops. I forgot you were there.

PAT: *90th percentile?!* You mean you seriously think there's a *1 in 10* chance that might happen?

ELIEZER: Ah, um...

STRANGER: Yes, he does. He wouldn't have considered it in exactly those words if I hadn't put it that way—not just because it's ridiculously specific, but because Eliezer Yudkowsky doesn't think in terms like that in advance of encountering the actual fact. He would consider it a "specific fantasy" that was threatening to drain away his emotional energy. But if it did happen, he would afterward say that he had achieved

an outcome such that around 10% of his probability mass "would have been" in outcomes like that one or better, though he would worry about being hindsight-biased.

PAT: I think a reasonable probability for an outcome like that would be more like 0.1%, and even that is being extremely generous!

ELIEZER: "Outside viewers" sure seem to tell me that a lot whenever I try to do anything interesting. I'm actually kind of surprised to hear you say that, though. I mean, my basic hypothesis for how the "outside view" thing operates is that it's an expression of incredulity that can be leveled against any target by cherry-picking a reference class that predicts failure. One then builds an inescapable epistemic trap around that reference class by talking about the Dunning-Kruger effect and the dangers of inside-viewing. But trying to write *Harry Potter* fanfiction, even unusually good *Harry Potter* fanfiction, should sound to most people like it's *not* high-status. I would expect people to react mainly to the part about the IMO gold medalist, even though the base rate for being an IMO gold medalist is higher than the base rate for authoring the most-reviewed *Harry Potter* fanfiction.

PAT: Have you ever even *tried* to write *Harry Potter* fanfiction before? Do you know *any* of the standard awards that help publicize the best *Harry Potter* fan works or any of the standard sites that recommend them? Do you have any idea what the vast majority of the audience for *Harry Potter* fanfiction *wants?* I mean, just the fact that you're publishing on FanFiction.Net is going to turn off a lot of people; the better stories tend to be hosted at ArchiveOfOurOwn.Org or on other, more specialized sites.

ELIEZER: Oh. I see. You *do* know about the pre-existing online *Harry Potter* fanfiction community, and you're involved in it. You actually *have* a pre-existing status hierarchy built up in your mind around *Harry Potter* fanfiction. So when the Masked Stranger talks about *Methods* becoming the most popular *Harry Potter* fanfiction ever, you really do hear that as an overreaching status-claim, and you do that thing that makes an arbitrary proposition sound very improbable using the "outside view."

PAT: I don't think the outside view, or reference class forecasting, can make arbitrary events sound very improbable. I think it makes events that *won't actually happen* sound very improbable. As for my prior acquaintance with the community—how is that supposed to devalue my opinions? I have domain expertise. I have some actual idea of how many thousands of authors, including some *very good* authors, are trying to write *Harry Potter* fanfiction, only one of whom can author the most-reviewed story. And I'll ask again, did you bother to acquire any idea of how this community actually works? Can you name a single annual award that's given out in the *Harry Potter* fanfiction community?

ELIEZER: Um... not off the top of my head.

PAT: Have you asked any of the existing top *Harry Potter* fanfiction authors to review your proposed plot, or your proposed story ideas? Like Nonjon, author of *A Black Comedy?* Or Sarah1281 or JBern or any of the other authors who have created multiple works widely acknowledged as excellent?

ELIEZER: I must honestly confess, although I've read those authors and liked their stories, that thought never even crossed my mind as a possible action.

PAT: So you haven't consulted anyone who knows more about *Harry Potter* fandom than you do.

ELIEZER: Nope.

PAT: You have not written any prior *Harry Potter* fanfiction—not even a short story.

ELIEZER: Correct.

PAT: You have made no previous effort to engage with the existing community of people who read or write *Harry Potter* fanfiction, or learn about existing gatekeepers on which the success of your story will depend.

ELIEZER: I've read some of the top previous *Harry Potter* fan works, since I enjoyed reading them. That, of course, is why the story idea popped into my head in the first place.

PAT: What would you think of somebody who'd read a few popular physics books and wanted to be the world's greatest physicist?

STRANGER: (*aside*) It appears to me that since the "outside view" as usually invoked is really about status hierarchy, signs of disrespecting the existing hierarchy will tend to provoke stronger reactions, and disrespectful-seeming claims that you can outperform some benchmark will be treated as much larger factors predicting failure than respectful-seeming claims that you can outperform an equivalent benchmark. It seems that physics crackpots feel relevantly analogous here because crackpots aren't just epistemically misguided—that would be tragicomic, but it wouldn't evoke the same feelings of contempt or disgust. What distinguishes physics crackpots is that they're epistemically misguided in ways that disrespect high-status people on an important hierarchy—physicists. This *feels* like a relevant reference class for understanding other apparent examples of disrespectfully claiming to be high-status, because the evoked feeling is similar even if the phenomena differ in other ways.

ELIEZER: If you want to be a great physicist, you have to find the *true* law of physics, which is already out there in the world and not known to you. This isn't something you can realistically achieve without working alongside other physicists, because you need an extraordinarily specific key to fit into this extraordinarily specific lock. In contrast, there are many possible books that would succeed over all past *Harry Potter* fanfiction, and you don't have to build a particle accelerator to figure out which one to write.

STRANGER: I notice that when you try to estimate the difficulty of becoming the greatest physicist ever, Eliezer, you try to figure out the difficulty of the corresponding cognitive problem. It doesn't seem to occur to you to focus on the *fame*.

PAT: Eliezer, you seem to be deliberately missing the point of what's wrong with reading a few physics books and then trying to become the world's greatest physicist. Don't you see that this error has the same *structure* as your *Harry Potter* pipe dream, even if the mistake's magnitude is greater? That a critic would say the same sort of things to them as I am saying to you? Yes, becoming the world's greatest physicist is much more difficult. But you're trying to do this lesser impossible task *in your off-hours* because you think it will be *easy.*

ELIEZER: In the success scenario the Masked Stranger described, I would invest more effort into later chapters because it would have proven to be worth it.

STRANGER: Hey, Pat? Did you know that Eliezer hasn't actually read the original *Harry Potter* books four through six, just watched the movies? And even after the book

starts to take off, he still won't get around to reading them.

PAT: (*incoherent noises*)

ELIEZER: Um... look, I read books one through three when they came out, and later I *tried* reading book four. The problem was, I'd already read so much *Harry Potter* fanfiction by then that I was used to thinking of the Potterverse as a place for grown-up stories, and this produced a state change in my brain, so when I tried to read *Harry Potter and the Goblet of Fire* it didn't feel right. But I've read enough fanfiction based in the Potterverse that I know the *universe* very well. I can tell you the name of Fleur Delacour's little sister. In fact, I've read an entire novel about Gabrielle Delacour. I just haven't read all the original books.

STRANGER: And when that's not good enough, Eliezer consults the *Harry Potter* Wikia to learn relevant facts from canon. So you see he has all the knowledge he thinks he needs.

PAT: (*more incoherent noises*)

ELIEZER: ...why did you tell Pat that, Masked Stranger?

STRANGER: Because Pat will think it's a tremendously relevant fact for predicting your failure. This illustrates a critical life lesson about the difference between making obeisances toward a field by reading works to demonstrate social respect, and trying to gather key knowledge from a field so you can advance it. The latter is necessary for success; the former is primarily important insofar as public relations with gatekeepers is important. I think that people who aren't status-blind have a harder time telling the difference.

PAT: It's true that I feel a certain sense of indignation—of, indeed, J. K. Rowling and the best existing *Harry Potter* fanfiction writers being actively disrespected—when you tell me that Eliezer *hasn't read all of the canon books* and that he thinks he'll make up for it by *consulting a wiki.*

ELIEZER: Well, if I can try to repair some of the public relations damage: If I thought I could write children's books as popular as J. K. Rowling's originals, I would be doing that instead. J. K. Rowling is now a billionaire, plus she taught my little sister to enjoy reading. People who trivialize that as "writing children's books" obviously have never tried to write anything themselves, let alone children's books. Writing good children's literature is *hard*—which is why *Methods* is going to be aimed at older readers. Contrary to the model you seem to be forming of me, I have a detailed model of my own limitations as well as my current capabilities, and I know that I am not currently a good enough author to write children's books.

PAT: I *can* imagine a state of affairs where I would estimate someone to have an excellent chance of writing the best *Harry Potter* fanfiction ever made, even after reading only the first three canon books—say, if Neil Gaiman tried it. (Though Neil Gaiman, I'm damned sure, just *would* read the original canon books.) Do you think you're as good as Neil Gaiman?

ELIEZER: I don't expect to ever have enough time to invest in writing to become as good as Neil Gaiman.

PAT: I've read your *Three Worlds Collide*, which I think is your best story, and I'm aware that it was mentioned favorably by a Hugo-award-winning author, Peter Watts.

But I don't think *Three Worlds Collide* is on the literary level of, say, the fanfiction *Always and Always Part 1: Backwards With Purpose.* So what feats of writing have you already performed that make you think your project has a 10% chance of becoming the most-reviewed *Harry Potter* fanfiction in existence?

ELIEZER: What you're currently doing is what I call "demanding to see my hero license." Roughly, I've declared my intention to try to do something that's in excess of what you think matches my current social standing, and you want me to show that I already have enough status to do it.

PAT: *Ad hominem*; you haven't answered my question. I don't see how, on the knowledge you presently have and on the evidence already available, you can possibly justify giving yourself a 10% probability here. But let me make sure, first, that we're using the same concepts. Is that "10%" supposed to be an actual well-calibrated probability?

ELIEZER: Yes, it is. If I interrogate my mind about betting odds, I think I'd take your money at 20:1—like, if you offered me $20 against $1 that the fanfiction wouldn't succeed—and I'd start feeling nervous about betting the other way at $4 against $1, where you'll pay out $4 if the fanfiction succeeds in exchange for $1 if it doesn't. Splitting the difference at somewhere near the geometric mean, we could call that 9:1 odds.

PAT: And do you think you're well-calibrated? Like, things you assign 9:1 odds should happen 9 out of 10 times?

ELIEZER: Yes, I think I could make 10 statements of this difficulty that I assign 90% probability, and be wrong on average about once. I haven't tested my calibration as extensively as some people in the rationalist community, but the last time I took a CFAR calibration-testing sheet with 10 items on them and tried to put 90% credibility intervals on them, I got exactly one true value outside my interval. Achieving okay calibration, with a bit of study and a bit of practice, is not anywhere near as surprising as outside-view types make it out to be.

STRANGER: (*aside*) Eliezer-2010 doesn't use PredictionBook as often as Gwern Branwen, doesn't play calibration party games as often as Anna Salamon and Carl Shulman, and didn't join Philip Tetlock's study on superprediction. But I did make bets whenever I had the opportunity, and still do; and I try to set numeric odds whenever I feel uncertain and know I'll find out the true value shortly.

I recently saw a cryptic set of statements on my refrigerator's whiteboard about a "boiler" and various strange numbers and diagrams, which greatly confused me for five seconds before I hypothesized that they were notes about Brienne's ongoing progress through the game *Myst*. Since I felt uncertain, but could find out the truth soon, I spent thirty seconds trying to tweak my exact probability estimate of these being notes for Brienne's game. I started with a 90% "first pass" probability that they were *Myst* notes, which felt obviously overconfident, so I adjusted that down to 80% or 4:1. Then I thought about how there might be unforeseen other compact explanations for the cryptic words on the whiteboard and adjusted down to 3:1. I then asked Brienne, and learned that it was in fact about her *Myst* game. I then did a thirty-second "update meditation" on whether perhaps it *wasn't* all that probable that there would be some other compact explanation for the cryptic writings; so maybe once the writings seemed explained away, I should have been less worried about unforeseen compact alternatives.

But I didn't meditate on it too long, because it was just one sample out of my life, and the point of experiences like that is that you have a lot of them, and update a little each time, and eventually the experience accumulates. Meditating on it as much as I'm currently doing by writing about it here would not be good practice in general. (Those of you who have a basic acquaintance with neural networks and the delta rule should recognize what I'm trying to get my brain to do here.) I feel guilty about not betting more systematically, but given my limited supply of spoons, this kind of informal and opportunistic but regular practice is about all that I'm likely to *actually* do, as opposed to feel guilty about not doing.

As I do my editing pass on this document, I more recently assigned 5:1 odds against two characters on *House of Cards* having sex, who did in fact have sex; and that provides a bigger poke of adjustment against overconfidence. (According to the delta rule, this was a bigger error.)

PAT: But there are studies showing that even after being warned about overconfidence, reading a study about overconfidence, and being allowed to practice a bit, overconfidence is *reduced* but not *eliminated*—right?

ELIEZER: On average across all subjects, overconfidence is reduced but not eliminated. That doesn't mean that in every individual subject*,* overconfidence is reduced but not eliminated.

PAT: What makes you think *you* can do better than average?

STRANGER: …

ELIEZER: What makes me think I could do better than average is that I practiced much more than those subjects, and I don't think the level of effort put in by the average subject, even a subject who's warned about overconfidence and given one practice session, is the limit of human possibility. And what makes me think I actually succeeded is that *I checked*. It's not like there's this "reference class" full of overconfident people who *hallucinate* practicing their calibration and *hallucinate* discovering that their credibility intervals have started being well-calibrated.

STRANGER: I offer some relevant information that I learned from Sarah Constantin's "Do Rational People Exist?": Stanovich and West (1997) found that 88% of study participants were systematically overconfident, which means that they couldn't demonstrate overconfidence for the remaining 12%. And this isn't too surprising; Stanovich and West (1998) note a number of other tests where around 10% of undergraduates fail to exhibit this or that bias.

ELIEZER: Right. So the question is whether I can, with some practice, make myself as non-overconfident as the top 10% of college undergrads. This… does not strike me as a particularly harrowing challenge. It does require effort. I have to consciously work to expand my credibility intervals past my first thought, and I expect that college students who outperform have to do the same. The potential to do better buys little of itself; you have to actually put in the effort. But when I think I've expanded my intervals enough, I stop.

ii. Success factors and belief sharing

PAT: So you *actually* think that you're well-calibrated in assigning 9:1 odds for *Methods* failing versus succeeding, to the extreme levels assigned by the Masked Stranger. Are you going to argue that I ought to widen my confidence intervals for how much success *Harry Potter and the Methods of Rationality* might enjoy, in order to avoid being overconfident myself?

ELIEZER: No. That feels equivalent to arguing that you shouldn't assign a 0.1% probability to *Methods* succeeding because 1,000:1 odds are too extreme. I was careful not to put it that way, because that isn't a valid argument form. That's the kind of thinking which leads to papers like Ord, Hillerbrand, and Sandberg's "Probing the Improbable," which I think are wrong. In general, if there are 500,000 fan works, only one of which can have the most reviews, then you can't pick out one of them at random and say that 500,000:1 is too extreme.

PAT: I'm glad you agree with this obvious point. And I'm not stupid; I recognize that your stories are better than average. 90% of *Harry Potter* fanfiction is crap by Sturgeon's Law, and 90% of the remaining 10% is going to be uninspired. That leaves maybe 5,000 fan works that you do need to seriously compete with. And I'll even say that if you're trying reasonably hard, you can end up in the top 10% of *that* pool. That leaves a 1-in-500 chance of your being the best *Harry Potter* author on fanfiction.net. We then need to factor in the other *Harry Potter* fanfiction sites, which have fewer works but much higher average quality. Let's say it works out to a 1-in-1,000 chance of yours being the best story ever, which I think is actually very generous of me, given that in a lot of ways you seem ridiculously unprepared for the task—um, are you all right, Masked Stranger?

STRANGER: Excuse me, please. I'm just distracted by the thought of a world where I could go on fanfiction.net and find 1,000 other stories as good as *Harry Potter and the Methods of Rationality.* I'm thinking of that world and trying not to cry. It's not that I can't *imagine* a world in which your modest-sounding Fermi estimate works correctly —it's just that the world you're describing looks so very different from this one.

ELIEZER: Pat, I can see where you're coming from, and I'm honestly not sure what I can say to you about it, in advance of being able to show you the book.

PAT: What about what *I* tried to say to *you?* Does it influence you *at all?* The method I used was rough, but I thought it was a very reasonable approach to getting a Fermi estimate, and if you disagree with the conclusion, I would like to know what further factors make your own Fermi estimate work out to 10%.

STRANGER: You underestimate the gap between how you two think. It wouldn't occur to Eliezer to even consider any one of the factors you named, while he was making his probability estimate of 10%.

ELIEZER: I have to admit that that's true.

PAT: Then what do *you* think are the most important factors in whether you'll succeed?

ELIEZER: Hm. Good question. I'd say... whether I can maintain my writing enthusiasm, whether I can write fast enough, whether I can produce a story that's really as good as I seem to be envisioning, whether I'll learn as I go and do better than I currently envision. Plus a large amount of uncertainty in how people will actually react to the work I have in my head if I can actually write it.

PAT: Okay, so that's five key factors. Let's estimate probabilities for each one. Suppose we grant that there's an 80% chance of your maintaining enthusiasm, a 50% chance that you'll write fast enough—though you've had trouble with that before; it took you fully a year to produce *Three Worlds Collide,* if I recall correctly. A 25% probability that you can successfully write down this incredible story that seems to be in your mind—I think this part almost always fails for authors, and is almost certainly the part that will fail for you, but we'll give it a one-quarter probability anyway, to be generous and steelman the whole argument. Then a 50% probability that you'll learn fast enough to not be torpedoed by the deficits you already know you have. Now even without saying *anything* about audience reactions (really, you're going to try to market cognitive science and formal epistemology to *Harry Potter* fans?), and even though I'm being very generous here, multiplying these probabilities together already gets us to the 5% level, which is less than the 10% you estimated—

STRANGER: Wrong.

PAT: … Wrong? What do you mean?

STRANGER: Let's consider the factors that might be involved in your above reasoning *not* being wrong. Let us first estimate the probability that any given English-language sentence will turn out to be true. Then, we have to consider the probability that a given argument supporting some conclusion will turn out to be free of fatal biases, the probability that someone who calls an argument "wrong" will be mistaken—

PAT: Eliezer, if you disagree with my conclusions, then what's wrong with my probabilities?

ELIEZER: Well, for a start: Whether I can maintain my writing speed is not *conditionally independent* of whether I maintain my enthusiasm. The audience reaction is not conditionally independent of whether I maintain my writing speed. Whether I'm learning things is not conditionally independent of whether I maintain my enthusiasm. Your attempt to multiply all those numbers together was gibberish as probability theory.

PAT: Okay, let's ask about the probability that you maintain writing speed, *given* that you maintain enthusiasm—

ELIEZER: Do you think that your numbers would have actually been that different, if that had been the question you'd initially asked? I'm pretty sure that if you'd thought to phrase the question as "the probability given that..." and hadn't first done it the other way, you would have elicited exactly the same probabilities from yourself, driven by the same balance of mental forces—picking something low that sounds reasonable, or something like that. And the problem of conditional dependence is far from the only reason I think "estimate these probabilities, which I shall multiply together" is just a rhetorical trick.

PAT: A rhetorical trick?

ELIEZER: By picking the right set of factors to "elicit," someone can easily make people's "answers" come out as low as desired. As an example, see van Boven and Epley's "The Unpacking Effect in Evaluative Judgments." The problem here is that people... how can I compactly phrase this... people tend to assign median-tending probabilities to any category you ask them about, so you can *very* strongly manipulate their probability distributions by picking the categories for which you "elicit" probabilities. Like, if you ask car mechanics about the possible causes of a car not

starting—experienced car mechanics, who see the real frequencies on a daily basis!—and you ask them to assign a probability to "electrical system failures" versus asking separately for "dead battery," "alternator problems," and "spark plugs," the *unpacked* categories get collectively assigned much greater total probability than the *packed* category.

PAT: But perhaps, when I'm unpacking things that can potentially go wrong, I'm just compensating for the planning fallacy and how people usually aren't pessimistic enough—

ELIEZER: Above all, the problem with your reasoning is that *the stated outcome does not need to be a perfect conjunction of those factors*. Not everything on your list has to go right simultaneously for the whole process to work. You have omitted other disjunctive pathways to the same end. In your universe, nobody ever tries harder or repairs something after it goes wrong! I have *never* yet seen an informal conjunctive breakdown of an allegedly low probability in which the final conclusion *actually required* every one of the premises. That's why I'm always careful to avoid the "I shall helpfully break down this proposition into a big conjunction and ask you to assign each term a probability" trick.

Its only real use, at least in my experience, is that it's a way to get people to feel like they've "assigned" probabilities while you manipulate the setup to make the conclusion have whatever probability you like—it doesn't have any role to play in honest conversation. Out of all the times I've seen it used, to support conclusions I endorse as well as ones I reject, I've never once seen it actually work as a way to better discover truth. I think it's bad epistemology that sticks around because it sounds sort of reasonable if you don't look too closely.

PAT: I was working with the factors *you* picked out as critical. Which *specific* parts of my estimate do you disagree with?

STRANGER: (*aside*) The multiple-stage fallacy is an *amazing* trick, by the way. You can ask people to think of key factors *themselves* and *still* manipulate them really easily into giving answers that imply a low final answer, because so long as people go on listing things and assigning them probabilities, the product is bound to keep getting lower. Once we realize that by continually multiplying out probabilities the product keeps getting lower, we have to apply some compensating factor internally so as to go on discriminating truth from falsehood.

You have effectively decided on the answer to most real-world questions as "no, *a priori*" by the time you get up to four factors, let alone ten. It may be wise to list out many possible failure scenarios and decide in advance how to handle them—that's Murphyjitsu—but if you start assigning "the probability that *X* will go wrong and not be handled, conditional on everything previous on the list having not gone wrong or having been successfully handled," then you'd better be willing to assign conditional probabilities near 1 for the kinds of projects that succeed sometimes—projects like *Methods*. Otherwise you're ruling out their success *a priori,* and the "elicitation" process is a sham.

Frankly, I don't think the underlying methodology is worth repairing. I don't think it's worth bothering to try to make a compensating adjustment toward higher probabilities. We just shouldn't try to do "conjunctive breakdowns" of a success probability where we make up lots and lots of failure factors that all get informal

probability assignments. I don't think you can get good estimates that way even if you try to compensate for the predictable bias.

ELIEZER: I did list my own key factors, and I do feel doubt about whether they'll work out. If I were really confident in them, I'd be assigning a higher probability than 10%. But besides having conditional dependencies, my factors also have disjunctive as well as conjunctive character; they don't all need to go right and stay right simultaneously. I could get far enough into *Methods* to acquire an audience, suddenly lose my writing speed, and *Methods* could still end up ultimately having a large impact.

PAT: So how do you manipulate those factors to arrive at an estimate of 10% probability of extreme success?

ELIEZER: I don't. That's not how I got my estimate. I found two brackets, 20:1 and 4:1, that I couldn't nudge further without feeling nervous about being overconfident in one direction or the other. In other words, the same way I generated my set of ten credibility intervals for CFAR's calibration test. Then I picked something in the logarithmic middle.

PAT: So you didn't even try to list out all the factors and then multiply them together?

ELIEZER: No.

PAT: Then where the heck does your 10% figure *ultimately* come from? Saying that you got two other cryptic numbers, 20:1 and 4:1, and picked something in the geometric middle, doesn't really answer the fundamental question.

STRANGER: I believe the technical term for the methodology is "pulling numbers out of your ass." It's important to practice calibrating your ass numbers on cases where you'll learn the correct answer shortly afterward. It's also important that you learn the limits of ass numbers, and don't make unrealistic demands on them by assigning multiple ass numbers to complicated conditional events.

ELIEZER: I'd say I reached the estimate… by thinking about the object-level problem? By using my domain knowledge? By having already thought a lot about the problem so as to load many relevant aspects into my mind, then consulting my mind's native-format probability judgment—with some prior practice at betting having already taught me a little about how to translate those native representations of uncertainty into 9:1 betting odds. I'm not sure what additional information you want here. If there's a way to produce genuinely, demonstrably superior judgments using some kind of break-it-down procedure, I haven't read about it in the literature and I haven't practiced using it yet. If you show me that you can produce 9-out-of-10 correct 90% credible intervals, and your intervals are narrower than my intervals, and you got them using a break-it-down procedure, I'm happy to hear about it.

PAT: So basically your 10% probability comes from inaccessible intuition.

ELIEZER: In this case? Yeah, pretty much. There's just too little I can say to you about why *Methods* might work, in advance of being able to show you what I have in mind.

PAT: If the reasoning inside your head is valid, why can't it be explained to me?

ELIEZER: Because I have private information, frankly. I know the book I'm trying to create.

PAT: Eliezer, I think one of the key insights you're ignoring here is that it should be a *clue* to you that you think you have *incommunicable* reasons for believing your *Methods of Rationality* project can succeed. Isn't being unable to convince other people of their prospects of success just the sort of experience that crackpots have when they set out to invent bad physics theories? Isn't this incommunicable intuition just the sort of justification that they would try to give?

ELIEZER: But the method you're using—the method you're calling "reference class forecasting"—is too demanding to actually detect whether someone will end up writing the world's most reviewed *Harry Potter* fanfiction, whether that's me or someone else. The fact that a modest critic can't be persuaded isn't Bayesian discrimination between things that will succeed and things that will fail; it isn't evidence.

PAT: On the contrary, I would think it very reasonable if Nonjon told me that he intended to write the most-reviewed *Harry Potter* fanfiction. Nonjon's *A Black Comedy* is widely acknowledged as one of the best stories in the genre, Nonjon is well-placed in influential reviewing and recommending communities—Nonjon might not be *certain* to write the most reviewed story ever, but he has legitimate cause to think that he is one of the top contenders for writing it.

STRANGER: It's interesting how your estimates of success probabilities can be well summarized by a single quantity that correlates very well with how respectable a person is within a subcommunity.

PAT: Additionally, even if my demands were unsatisfiable, that wouldn't necessarily imply a hole in my reasoning. Nobody who buys a lottery ticket can possibly satisfy me that they have good reason to believe they'll win, even the person who *does* win. But that doesn't mean I'm wrong in assigning a low success probability to people who buy lottery tickets.

Nonjon may legitimately have a 1-in-10 lottery ticket. Neil Gaiman might have 2-in-3. Yours, as I've said, is probably more like 1-in-1,000, and it's only that high owing to your having already demonstrated some good writing abilities. I'm not even penalizing you for the fact that your plan of offering explicitly rational characters to the *Harry Potter* fandom sounds very unlike existing top stories. I might be unduly influenced by the fact that I like your previous writing. But your claim to have *incommunicable* advance knowledge that your lottery ticket will do better than this by a factor of 100 seems very suspicious to me. Valid evidence should be communicable between people.

STRANGER: "I believe myself to be writing a book on economic theory which will largely revolutionize—not I suppose, at once but in the course of the next ten years—the way the world thinks about its economic problems. I can't expect you, or anyone else, to believe this at the present stage. But for myself I don't merely hope what I say,—in my own mind, I'm quite sure." Lottery winner John Maynard Keynes to George Bernard Shaw, while writing *The General Theory of Employment, Interest and Money*.

ELIEZER: Come to think of it, if I *do* succeed with *Methods*, Pat, you yourself could end up in an incommunicable epistemic state relative to someone who only heard about me later through my story. Someone like that might *suspect* that I'm not a purely random lottery ticket winner, but they won't have as much evidence to that effect as you. It's a pretty interesting and fundamental epistemological issue.

PAT: I disagree. If you have valid introspective evidence, then talk to me about your state of mind. On *my* view, you shouldn't end up in a situation where you update differently on what your evidence "feels like to you" than what your evidence "sounds like to other people"; both you and other people should just do the second update.

STRANGER: No, in this scenario, in the presence of other suspected biases, two human beings really can end up in incommunicable epistemic states. *You* would know that "Eliezer wins" had genuinely been singled out in advance as a distinguished outcome, but the second person would have to assess this supposedly distinguished outcome with the benefit of hindsight, and they may legitimately never trust their hindsight enough to end up in the same mental state as you.

You're right, Pat, that completely unbiased agents who lack truly foundational disagreements on priors should never end up in this situation. But humans can end up in it very easily, it seems to me. Advance predictions have special authority in science for a reason: hindsight bias makes it hard to ever reach the same confidence in a prediction that you only hear about after the fact.

PAT: Are you really suggesting that the prevalence of cognitive bias means you should be *more* confident that your own reasoning is correct? My epistemology seems to be much more straightforward than yours on these matters. Applying the "valid evidence should be communicable" rule to this case: A hypothetical person who saw Eliezer Yudkowsky write the *Less Wrong* Sequences, heard him mention that he assigned a non-tiny probability to succeeding in his *Methods* ambitions, and then saw him succeed at *Methods* should just realize what an external observer would say to them about that. And what they'd say is: you just happened to be the lucky or unlucky relatives of a lottery ticket buyer who claimed in advance to have psychic powers, and then happened to win.

ELIEZER: This sounds a lot like a difficulty I once sketched out for the "method of imaginary updates." Human beings aren't logically omniscient, so we can't be sure we've reasoned correctly about prior odds. In advance of seeing *Methods* succeed, I can see why you'd *say* that, on your worldview, if it did happen then it would just be a 1000:1 lottery ticket winning. But if that *actually happened*, then instead of saying, "Oh my gosh, a 1000:1 event just occurred," you ought to consider instead that the method you used to assign prior probabilities was flawed. This is not true about a lottery ticket, because we're extremely sure about how to assign prior probabilities in that case—and by the same token, in real life neither of us will *actually* see our friends winning the lottery.

PAT: I agree that if it actually happens, I would reconsider your previous arguments rather than insisting that I was correct about prior odds. I'm happy to concede this point because I am very, very confident that it won't actually happen. The argument against your success in *Harry Potter* fanfiction seems to me about as strong as any argument the outside-view perspective might make.

STRANGER: Oh, we aren't disputing that.

PAT: You aren't?

STRANGER: That's the whole point, from my perspective. If modest epistemology sounds persuasive to you, then it's trivial to invent a crushing argument against any project that involves doing something important that hasn't been done in the past. Any project that's trying to exceed any variety of civilizational inadequacy is going to be ruled out.

PAT: Look. You *cannot* just waltz into a field and become its leading figure on your first try. Modest epistemology is just *right* about that. You are not *supposed* to be able to succeed when the odds against you are like those I have described. Maybe out of a million contenders, someone will succeed by luck when the modest would have predicted their failure, but if we're batting 999,999 out of 1,000,000 I say we're doing pretty well. Unless, of course, Eliezer would claim that the project of writing this new *Harry Potter* fanfiction is so important that a 0.0001% chance of success is still worth it—

ELIEZER: I never say that. Ever. If I ever say that you can just shoot me.

PAT: Then *why* are you not responding to the very clear, very standard, very obvious reasons I have laid out to think that you cannot do this? I mean, seriously, what is going through your head right now?

ELIEZER: A helpless feeling of being unable to communicate.

STRANGER: Grim amusement.

PAT: Then I'm sorry, Mr. Eliezer Yudkowsky, but it seems to me that you are being irrational. You aren't even trying to hide it very hard.

ELIEZER: (*sighing*) I can imagine why it would look that way to you. I know how to communicate some of the thought patterns and styles that I think have served me well, that I think generate good predictions and policies. The other patterns leave me with this helpless feeling of knowing but being unable to speak. This conversation has entered a dependency on the part that I know but don't know how to say.

PAT: Why should I believe that?

ELIEZER: If you think the part I *did* figure out how to say was impressive enough. That was hidden purpose #7 of the *Less Wrong* Sequences—to provide an earnest-token of all the techniques I *couldn't* show. All I can tell you is that everything you're so busy worrying about is not the correct thing for me to be thinking about. That your entire approach to the problem is wrong. It is not just that your arguments are wrong. It is that they are about the wrong subject matter.

PAT: Then what's the right subject matter?

ELIEZER: That's what I'm having trouble saying. I can say that you ought to discard all thoughts from your mind about competing with others. The others who've come before you are like probes, flashes of sound, pingbacks that give you an incomplete sonar of your problem's difficulty. Sometimes you can swim past the parts of the problem that tangled up other people and enter a new part of the ocean. Which doesn't actually mean you'll succeed; all it means is that you'll have very little information about which parts are difficult. There often isn't actually any need to think at all about the intrinsic qualities of your competition—like how smart or motivated or well-paid they are—because their work is laid out in front of you and you can just look at the quality of the work.

PAT: Like somebody who predicts hyperinflation, saying all the while that they're free to disregard conventional economists because of how those idiot economists think you can triple the money supply without getting inflation?

ELIEZER: I don't really *know* what goes through someone else's mind when that happens to them. But I don't think that telling them to be more modest is a fix. Telling somebody to shut up and respect academics is not a generally valid line of argumentation because it doesn't distinguish mainstream economics (which has relatively high scholarly standards) from mainstream nutrition science (which has relatively low scholarly standards). I'm not sure there *is* any robust way out except by understanding economics for yourself, and to the extent that's true, I ought to advise our hypothetical ill-informed contrarian to read a lot of economics blogs and try to follow the arguments, or better yet read an economics textbook. I don't think that people sitting around and anxiously questioning themselves and wondering whether they're too audacious is a route out of that particular hole—let alone the hole on the other side of the fence.

PAT: So your meta-level epistemology is to remain as ultimately inaccessible to me as your object-level estimates.

ELIEZER: I can understand why you're skeptical.

PAT: I somehow doubt that you could pass an Ideological Turing Test on my point of view.

STRANGER: (*smiling*) Oh, I think I'd do pretty well at your ITT.

ELIEZER: Pat, I understand where your estimates are coming from, and I'm sure that your advice is truly meant to be helpful to me. But I also see that advice as an expression of a kind of *anxiety* which is not at all like the things I need to actually think about in order to produce good fiction. It's a wasted motion, a thought which predictably will not have helped in retrospect if I succeed. How good I am relative to other people is just not something I should spend lots of time obsessing about in order to make *Methods* be what I want it to be. So my thoughts just don't go there.

PAT: This notion, "that thought will predictably not have helped in retrospect if I succeed," seems very strange to me. It helps precisely because we can avoiding wasting our effort on projects which are unlikely to succeed.

STRANGER: Sounds very reasonable. All I can say in response is: try doing it my way for a day, and see what happens. No thoughts that predictably won't have been helpful in retrospect, in the case that you succeed at whatever you're currently trying to do. You might learn something from the experience.

ELIEZER: The thing is, Pat... even answering your objections and defending myself from your variety of criticism trains what look to me like unhealthy habits of thought. You're relentlessly focused on *me* and my psychology, and if I engage with your arguments and try to defend myself, I have to focus on *myself* instead of my book. Which gives me that much less attention to spend on sketching out what Professor Quirrell will do in his first Defense lesson. Worse, I have to defend my decisions, which can make them harder to change later.

STRANGER: Consider how much more difficult it will be for Eliezer to swerve and drop his other project, *The Art of Rationality*, if it fails after he has a number of (real or internal) conversation like this—conversations where he has to defend all the reasons why it's okay for him to think that he *might* write a nonfiction bestseller about rationality. This is why it's important to be able to *casually* invoke civilizational inadequacy. It's important that people be allowed to try ambitious things *without* feeling like they need to make a great production out of defending their hero license.

ELIEZER: Right. And... the mental motions involved in worrying what a critic might think and trying to come up with defenses or concessions are *different* from the mental motions involved in being curious about some question, trying to learn the answer, and coming up with tests; and it's different from how I think when I'm working on a problem in the world. The thing I should be thinking about is just the work itself.

PAT: If you were just trying to write okay *Harry Potter* fanfiction for fun, I might agree with you. But you say you can produce the *best* fanfiction. That's a whole different ball game—

ELIEZER: No! The perspective I'm trying to show you, the way it works in the inside of my head, is that trying to write good fanfiction, and the best fanfiction, are *not* different ball games. There's an object level, and you try to optimize it. You have an estimate of how well you can optimize it. That's all there ever is.

iii. Social heuristics and problem importance, tractability, and neglectedness

PAT: A funny thought has just occurred to me. That thing where you're trying to work out the theory of Friendly AI—

ELIEZER: Let me guess. You don't think I can do that either.

PAT: Well, I don't think you can save the world, of course! (*laughs*) This isn't a science fiction book. But I do see how you can reasonably hope to make an important contribution to the theory of Friendly AI that ends up being useful to whatever group ends up developing general AI. What's interesting to note here is that the scenario the Masked Stranger described, the class of successes you assigned 10% aggregate probability, is actually *harder* to achieve than that.

STRANGER: (*smiling*) It really, really, *really* isn't.

I'll mention as an aside that talk of "Friendly" AI has been going out of style where I'm from. We've started talking instead in terms of "aligning smarter-than-human AI with operators' goals," mostly because "AI alignment" smacks less of anthropomorphism than "friendliness."

ELIEZER: Alignment? Okay, I can work with that. But Pat, you've said something I didn't expect you to say and gone outside my current vision of your Ideological Turing Test. Please continue.

PAT: Okay. Contrary to what you think, my words are *not* fully general counterarguments that I launch against just anything I intuitively dislike. They are based on specific, visible, third-party-assessable factors that make assertions believable or unbelievable. If we leave aside inaccessible intuitions and just look at third-party-visible factors, then it is very clear that there's a huge community of writers who are *explicitly* trying to create *Harry Potter* fanfiction. This community is far larger and has far more activity—by every objective, third-party metric—than the community working on issues related to alignment or friendliness or whatever. Being the best writer in a much larger community is much more improbable than your

making a significant contribution to AI alignment when almost nobody else is working on *that* problem.

ELIEZER: The relative size of existing communities that you've just described is not a fact that I regard as important for assessing the relative difficulty of "making a key contribution to AI alignment" versus "getting *Methods* to the level described by the Masked Stranger." The number of competing fanfiction authors would be informative to me *if* I hadn't already checked out the *Harry Potter* fan works with the best reputations. If I can see how strong the competition is with my own eyes, then that screens off information about the size of the community from my perspective.

PAT: But surely the size of the community should give you some pause regarding whether you should *trust* your felt intuition that you could write something better than the product of so many other authors.

STRANGER: See, that meta-reasoning right there? That's the part I think is going to completely compromise how people think about the world if they try to reason that way.

ELIEZER: Would you ask a juggler, in the middle of juggling, to suddenly start worrying about whether she's in a reference class of people who merely think that they're good at catching balls? It's all just... wasted motion.

STRANGER: Social anxiety and overactive scrupulosity.

ELIEZER: Not what brains look like when they're thinking productively.

PAT: You've been claiming that the outside view is a fully general counterargument against any claim that someone with relatively low status will do anything important. I'm explaining to you why the method of trusting externally visible metrics and things that third parties can be convinced of says that you *might* make important contributions to AI alignment where nobody else is trying, but that you *won't* write the most reviewed *Harry Potter* fanfiction where thousands of other authors are competing with you.

(*A* WANDERING BYSTANDER *suddenly steps up to the group, interjecting.*)

BYSTANDER: Okay, no. I just can't hold my tongue anymore.

PAT: Huh? Who are you?

BYSTANDER: I am the *true* voice of modesty and the outside view!

I've been overhearing your conversation, and I've got to say—there's *no way* it's easier to make an important contribution to AI alignment than it is to write popular fanfiction.

ELIEZER: … That's true enough, but who…?

BYSTANDER: The name's Maude Stevens.

PAT: Well, it's nice to make your acquaintance, Maude. I am always eager to hear about my mistakes, even from people with suspiciously relevant background information who randomly walk up to me in parks. What is my error on this occasion?

MAUDE: All three of you have been taking for granted that if people don't talk about "alignment" or "friendliness," then their work isn't relevant. But those are just words. When we take into account machine ethicists working on real-world trolley dilemmas, economists working on technological unemployment, computer scientists working on Asimovian agents, and so on, the field of competitors all trying to make progress on these issues becomes much, much larger.

PAT: What? Is that true, Eliezer?

ELIEZER: Not to my knowledge—unless Maude is here from the NSA to tell me about some *very* interesting behind-closed-doors research. The examples Maude listed aren't addressing the technical issues I've been calling "friendliness." Progress on those problems doesn't help you with specifying preferences that you can reasonably expect to produce good outcomes even when the system is smarter than you and searching a much wider space of strategies than you can consider or check yourself. Or designing systems that are stable under self-modification, so that good properties of a seed AI are preserved as the agent gets smarter.

MAUDE: And your claim is that no one else in the world is smart enough to notice any of this?

ELIEZER: No, that's not what I'm saying. Concerns like "how do we specify correct goals for par-human AI?" and "what happens when AI gets smart enough to automate AI research itself?" have been around for a long time, sort of just hanging out and not visibly shifting research priorities. So it's not that the community of people who have ever thought about superintelligence is small; and it's not that there are no ongoing lines of work on robustness, transparency, or security in narrow AI systems that will incidentally make it easier to align smarter-than-human AI. But the community of people who go into work every day and make decisions about what technical problems to tackle based on any extended thinking related to superintelligent AI is very small.

MAUDE: What I'm saying is that you're jumping ahead and trying to solve the far end of the problem before the field is ready to focus efforts there. The current work may not all bear directly on superintelligence, but we should expect all the significant progress on AI alignment to be produced by the intellectual heirs of the people presently working on topics like drone warfare and unemployment.

PAT: (*cautiously*) I mean, if what Eliezer says is true—and I *do* think that Eliezer is honest, if often, by my standards, slightly crazy—then the state of the field in 2010 is just like it looks naively. There aren't many people working on topics related to smarter-than-human AI, and Eliezer's group and the Oxford Future of Humanity Institute are the only ones with a reasonable claim to be working on AI alignment. If Eliezer says that the problems of crafting a smarter-than-human AI to not kill everyone are not of a type with current machine ethics work, then I can buy that as plausible, though I'd want to hear others' views on the issue before reaching a firm conclusion.

MAUDE: But Eliezer's field of competition is far wider than just the people writing ethics papers. Anyone working in machine learning, or indeed in any branch of computer science, might end up contributing to AI alignment.

ELIEZER: Um, that would certainly be great news to hear. The win state here *is* just "the problem gets solved"—

PAT: Wait a second. I think you're leaving the realm of what's third-party objectively verifiable, Maude. That's like saying that Eliezer has to compete with Stephen King because Stephen King could in principle decide to start writing *Harry Potter* fanfiction. If all these other people in AI are *not* working on the particular problems Eliezer is working on, whereas the broad community of *Harry Potter* fanfiction writers is competing *directly* with Eliezer on fiction-writing, then any reasonable third party should agree that the outside view counterargument applies very strongly to the second case, and much more weakly (if at all) to the first.

MAUDE: So now fanfiction is supposed to be harder than saving the world? Seriously? Just no.

ELIEZER: Pat, while I disagree with Maude's arguments, she does have the advantage of rationalizing a true conclusion rather than a false conclusion. AI alignment *is* harder.

PAT: I'm not expecting you to solve the whole thing. But making a significant contribution to a sufficiently specialized corner of academia that very few other people are explicitly working on should be easier than becoming the *single most successful figure* in a field that lots of other people are working in.

MAUDE: This is ridiculous. Fanfiction writers are simply not the same kind of competition as machine learning experts and professors at leading universities, any of whom could end up making far more impressive contributions to the cutting edge in AGI research.

ELIEZER: Um, advancing AGI research might be *impressive*, but unless it's AGI *alignment* it's—

PAT: Have you ever *tried* to write fiction yourself? Try it. You'll find it's a heck of a lot harder than you seem to imagine. Being good at math does *not* qualify you to waltz in and—

(*The Masked Stranger raises his hand and snaps his fingers. All time stops. Then the Masked Stranger looks over at Eliezer-2010 expectantly.*)

ELIEZER: Um... Masked Stranger... do you have any idea what's going on here?

STRANGER: Yes.

ELIEZER: Thank you for that concise and informative reply. Would you please *explain* what's going on here?

STRANGER: Pat is thoroughly acquainted with the status hierarchy of the established community of *Harry Potter* fanfiction authors, which has its own rituals, prizes, politics, and so on. But Pat, for the sake of literary hypothesis, lacks an instinctive sense that it's audacious to try to contribute work to AI alignment. If we interrogated Pat, we'd probably find that Pat believes that alignment is cool but not astronomically important, or that there are many other existential risks of equal stature. If Pat

believed that long-term civilizational outcomes depended mostly on solving the alignment problem, as you do, then he would probably assign the problem more instinctive prestige—*holding constant* everything Pat knows about the object-level problem and how many people are working on it, but raising the problem's felt status.

Maude, meanwhile, is the reverse: not acquainted with the political minutiae and status dynamics of *Harry Potter* fans, but very sensitive to the importance of the alignment problem. So to Maude, it's intuitively obvious that making technical progress on AI alignment requires a much more impressive hero license than writing the world's leading *Harry Potter* fanfiction. Pat doesn't see it that way.

ELIEZER: But ideas in AI alignment have to be formalized; and the formalism needs to satisfy many different requirements simultaneously, without much room for error. It's a very abstract, very highly *constrained* task because it has to put an informal problem into the *right* formal structure. When writing fiction, yes, I have to juggle things like plot and character and tension and humor, but that's all still a much less constrained cognitive problem—

STRANGER: That kind of consideration isn't likely to enter Pat or Maude's minds.

ELIEZER: Does it matter that I intend to put far more effort into my research than into fiction-writing? If *Methods* doesn't work the first time, I'll just give up.

STRANGER: Sorry. Whether or not you're allowed to do high-status things *can't* depend on how much effort you say you intend to put in. Because "anyone could say that." And then you couldn't slap down pretenders—which is terrible.

ELIEZER: ..... Is there some kind of organizing principle that makes all of this make sense?

STRANGER: I think the key concepts you need are *civilizational inadequacy* and *status hierarchy maintenance.*

ELIEZER: Enlighten me.

STRANGER: You know how Pat ended up calculating that there ought to be 1,000 works of *Harry Potter* fanfiction as good as *Methods*? And you know how I got all weepy visualizing that world? Imagine Maude as making a similar mistake. There's a world in which some scruffy outsider like you *wouldn't* be able to estimate a significant chance of making a major contribution to AI alignment, let alone help found the field, because people had been trying to do serious technical work on it since the 1960s, and were putting substantial thought, ingenuity, and care into making sure they were working on the right problems and using solid methodologies. Functional decision theory was developed in 1971, two years after Robert Nozick's publication of "Newcomb's Problem and Two Principles of Choice." Everyone expects humane values to have high Kolmogorov complexity. Everyone understands why, if you program an expected utility maximizer with utility function **U** and what you really meant is *V*, the **U**-maximizer has a convergent instrumental incentive to deceive you into believing that it is a *V*-maximizer. Nobody assumes you can "just pull the plug" on something much smarter than you are. And the world's other large-scale activities and institutions all scale up similarly in competence.

We could call this the Adequate World, and contrast it to the way things actually are. The Adequate World has a property that we could call *inexploitability*; or inexploitability-by-Eliezer. We can compare it to how you can't predict a 5% change in

Microsoft's stock price over the next six months—take that property of S&P 500 stocks, and scale it up to a whole planet whose experts you can't surpass, where *you* can't find any *knowable* mistake. They still make mistakes in the Adequate World, because they're not perfect. But they're smarter and nicer at the group level than Eliezer Yudkowsky, so you can't know which things are epistemic or moral mistakes, just like you can't know whether Microsoft's equity price is mistaken on the up-side or low-side on average.

ELIEZER: Okay... I can see how Maude's *conclusion* would make sense in the Adequate World. But how does Maude reconcile the arguments that reach that conclusion with the vastly different world we actually live in? It's not like Maude can say, "Look, it's obviously already being handled!" because it obviously *isn't*.

STRANGER: Suppose that you have an instinct to regulate status claims, to make sure nobody gets more status than they deserve.

ELIEZER: Okay...

STRANGER: This gives rise to the behavior you've been calling "hero licensing." Your current model is that people have read too many novels in which the protagonist is born under the sign of a supernova and carries a legendary sword, and they don't realize real life is not like that. Or they associate the deeds of Einstein with the prestige that Einstein has *now*, not realizing that prior to 1905, Einstein had no visible aura of destiny.

ELIEZER: Right.

STRANGER: Wrong. Your model of heroic status is that it ought to be a reward for heroic service to the tribe. You think that while of course we should discourage people from claiming this heroic status without having yet served the tribe, no one should find it intuitively objectionable to merely *try* to serve the tribe, as long as they're careful to disclaim that they haven't *yet* served it and don't claim that they already deserve the relevant status boost.

ELIEZER: ... this is wrong?

STRANGER: It's fine for "status-blind" people like you, but it isn't how the standard-issue status emotions work. Simply put, there's a level of status you need in order to reach up for a given higher level of status; and this is a relatively basic feeling for most people, not something that's trained into them.

ELIEZER: But before 1905, Einstein was a patent examiner. He didn't even get a PhD until 1905. I mean, Einstein wasn't a *typical* patent examiner and he no doubt knew that himself, but someone on the outside looking at just his CV—

STRANGER: We aren't talking about an epistemic prediction here. This is just a fact about how human status instincts work. Having a certain probability of writing the most popular *Harry Potter* fanfiction in the future comes with a certain amount of status in Pat's eyes. Having a certain probability of making important progress on the AI alignment problem in the future comes with a certain amount of status in Maude's eyes. Since your current status in the relevant hierarchy seems much lower than that, you aren't allowed to endorse the relevant probability assignments or act as though you think they're correct. You are not allowed to just try it and see what happens, since that already implies that you think the probability is non-tiny. The very act of

affiliating yourself with the possibility is status-overreaching, requiring a slapdown. Otherwise any old person will be allowed to claim too much status—which is terrible.

ELIEZER: Okay. But how do we get from there to delusions of civilizational adequacy?

STRANGER: Backward chaining of rationalizations, perhaps mixed with some amount of just-world and status-quo bias. An economist would say "What?" if you presented an argument saying you ought to be able to double your money every year by buying and selling Microsoft stock in some simple pattern. The economist would then, quite reasonably, initiate a mental search to try to come up with some way that your algorithm doesn't do what you thought it did, a hidden risk it contained, a way to preserve the idea of an inexploitable market in equities.

Pat tries to preserve the idea of an inexploitable-by-Eliezer market in fanfiction (since on a gut level it feels to him like you're too low-status to be able to exploit the market), and comes up with the idea that there are a thousand other people who are writing equally good *Harry Potter* fanfiction. The result is that Pat hypothesizes a world that is *adequate* in the relevant respect. Writers' efforts are cheaply converted into stories so popular that it's just about humanly impossible to foreseeably write a more popular story; and the world's adequacy in other regards ensures that any outsiders who *do* have a shot at outperforming the market, like Neil Gaiman, will already be rich in money, esteem, etc.

And the phenomenon generalizes. If someone believes that you don't have enough status to make better predictions than the European Central Bank, they'll have to believe that the European Central Bank is reasonably good at its job. Traditional economics doesn't say that the European Central Bank has to be good at its job—an economist would tell you to look at incentives, and that the decisionmakers don't get paid huge bonuses if Europe's economy does better. For the status order to be preserved, however, it can't be possible for Eliezer to outsmart the European Central Bank. For the world's status order to be unchallengeable, it has to be right and wise; for it to be right and wise, it has to be inexploitable. A gut-level appreciation of civilizational inadequacy is a powerful tool for dispelling mirages like hero licensing and modest epistemology, because when modest epistemology backward-chains its rationalizations for why you can't achieve big things, it ends up asserting adequacy.

ELIEZER: Civilization could be inexploitable in these areas without being adequate, though; and it sounds like you're saying that Pat and Maude mainly care about inexploitability.

STRANGER: You could have a world where poor incentives result in alignment research visibly being neglected, but where there's no realistic way for well-informed and motivated individuals to strategically avoid those incentives without being outcompeted in some other indispensable resource. You could also have a world that's inexploitable to you but exploitable to many other people. However, asserting adequacy reaffirms the relevant status hierarchy in a much stronger and more airtight way. The notion of an Adequate World more closely matches the intuitive sense that the world's most respectable and authoritative people are just untouchable—too well-organized, well-informed, and well-intentioned for just anybody to spot Moloch's handiwork, whether or not they can do anything about it. And affirming adequacy in a way that sounds vaguely plausible generally requires less detailed knowledge of microeconomics, of the individuals trying to exploit the market, and of the specific problems they're trying to solve than is the case for appeals to inexploitable inadequacy.

Civilizational inadequacy is the basic reason why the world as a whole isn't inexploitable in the fashion of short-term equity price changes. The modest view, roughly, is that the world *is* inexploitable as far as you can predict, because you can never *knowably* know better than the experts.

ELIEZER: I... sort of get it? I still don't understand Maude's actual thought process here.

STRANGER: Let's watch, then.

(*The Masked Stranger raises his hands and snaps his fingers again, restarting time.*)

PAT: —take over literature because mere fiction writers are stupid.

MAUDE: My good fellow, please take a moment to consider what you're proposing. If the AI alignment problem were really as important as Eliezer claims, would he really be one of the only people working on it?

PAT: Well, it sure looks like he is.

MAUDE: Then the problem can't be as important as he claims. The alternative is that a lone crank has identified an important issue that he and very few others are working on; and that means everyone else in his field is an idiot. Who does Eliezer think he is, to defy the academic consensus to the effect that AI alignment isn't an interesting idea worth working on?

PAT: I mean, there are all sorts of barriers I could imagine a typical academic running into if they wanted to work on AI alignment. Maybe it's just hard to get academic grants for this kind of work.

MAUDE: If it's hard to get grants, then that's because the grant-makers correctly recognize that this isn't a priority problem.

PAT: So now the *state of academic funding* is said to be so wise that people can't find neglected research opportunities?

STRANGER: What person with grant-making power gets paid less in the worlds where alignment is important and yet neglected? If no one loses their bonuses or incurs any other perceptible cost, then you're done. There's no mystery here.

MAUDE: All of the evidence is perfectly consistent with the hypothesis that there are no academic grants on offer because the grantmakers have made a thoughtful and informed decision that this is a pseudo-problem.

ELIEZER: I appreciate Pat's defense, but I think I can better speak to this. Issues like intelligence explosion and the idea that there's an important problem to be solved in AI goal systems, as I mentioned earlier, aren't original to me. They're reasonably widely known, and people at all levels of seniority are often happy to talk about it face-to-face, though there's disagreement about the magnitude of the risk and about what kinds of efforts are likeliest to be useful for addressing it. You can find it discussed in the most commonly used undergrad textbook in AI, *Artificial Intelligence:*

A Modern Approach. You can't claim that there's a consensus among researchers that this is not an important problem.

MAUDE: Then the grantmakers probably carefully looked into the problem and determined that the best way to promote humanity's long-term welfare is to advance the field of AI in other ways, and only work on alignment once we reach some particular capabilities threshold. At that point, in all likelihood, funders plan to coordinate to launch a major field-wide research effort on alignment.

ELIEZER: How, exactly, could they reach a conclusion like that without studying the problem in any visible way? If the entire grantmaking community was able to arrive at a consensus to that effect, then where are the papers and analyses they used to reach their conclusion? What are the arguments? You sound like you're talking about a *silent conspiracy* of competent grantmakers at a hundred different organizations, who have in some way collectively developed or gained access to a literature of strategic and technical research that Nick Bostrom and I have never heard about, establishing that the present-day research problems that look relevant and tractable aren't so promising, and that capabilities will develop in a specific known direction at a particular rate that lends itself to late coordinated intervention.

Are you saying that despite all the researchers in the field casually discussing self-improving AI and Asimov Laws over coffee, there's some hidden clever reason why studying this problem isn't a good idea, which the grantmakers all arrived at in unison without leaving a paper trail about their decision-making process? I just... There are so many well-known and perfectly normal dysfunctions of grantmaking machinery and the academic incentive structure that allow alignment to be a critical problem without there necessarily being a huge academic rush to work on it. Instead you're postulating a massive global conspiracy of hidden competence grounded in secret analyses and arguments. Why would you possibly go there?

MAUDE: Because otherwise—

(*The Stranger snaps his fingers again.*)

STRANGER: Okay, Eliezer-2010, go ahead and answer. Why is Maude going there?

ELIEZER: Because... to prevent relatively unimpressive or unauthoritative-looking people from affiliating with important problems, from Maude's perspective there can't be knowably low-hanging research fruit. If there were knowably important problems that the grantmaking machinery and academic reward system had left untouched, then somebody like me could knowably be working on them. If there were a problem with the grantmakers, or a problem with academic incentives, at least of the kind that someone like me could identify, then it might be possible for someone unimportant like me to know that an important problem was not being worked on. The alleged state of academia and indeed the whole world has to backward chain to avoid there being low-hanging research fruit.

First Maude tried to argue that the problem is already well-covered by researchers in the field, as it would be in the Adequate World you described. When that position became difficult to defend, she switched to arguing that authoritative analysts have looked into the problem and collectively determined it's a pseudo-problem. When that

became difficult to defend, she switched to arguing that authoritative analysts have looked into the problem and collectively devised a better strategy involving delaying alignment research temporarily.

STRANGER: Very different hypotheses that share this property: they allow there to be something like an efficient market in high-value research, where individuals and groups that have high status in the standard academic system can't end up visibly dropping the ball.

Perhaps Maude's next proposal will be that top researchers have determined that the problem is easy. Perhaps there's a hidden consensus that AGI is centuries away. In my experience, people like Maude can be boundlessly inventive. There's always something.

ELIEZER: But why go to such lengths? No real economist would tell us to expect an efficient market here.

STRANGER: Sure, says Maude, the system isn't perfect. But, she continues, neither are we perfect. All the grantmakers and tenure-granters are in an equivalent position to us, and doing their own part to actively try to compensate for any biases in the system they think they can see.

ELIEZER: But that's visibly contradicted both by observation and by the economic theory of incentives.

STRANGER: Yes. But at the same time, it has to be assumed true. Because while experts can be wrong, we can also be wrong, right? Maybe *we're* the ones with bad systemic incentives and only short-term rewards.

ELIEZER: But being inside a system with badly designed incentives is *not* the same as being unable to discern the truth of... oh, never mind.

This has all been very educational, Masked Stranger. Thanks.

STRANGER: Thanks for what, Eliezer? Showing you a problem isn't much of a service if there's nothing you can do to fix it. You're no better off than you were in the original timeline.

ELIEZER: It still feels better to have some idea of what's going on.

STRANGER: That, too, is a trap, as we're both aware. If you need an elaborate theory to justify seeing the obvious, it will only become more elaborate and distracting as time goes on and you try harder and harder to reassure yourself. It's much better to just take things at face value, without needing a huge argument to do so. If you must ignore someone's advice, it's better not to make up big elaborate reasons why you're licensed to ignore it; that makes it easier to change your mind and take the advice later, if you happen to feel like it.

ELIEZER: True. Then why are you even saying these things to me?

STRANGER: I'm not. You never were the one to whom I was speaking, this whole time. That is the last lesson, that I didn't ever say these things to myself.

(*The Stranger turns upon his own heel three times, and was never there.*)

Sunset at Noon

A meandering series of vignettes.

I have a sense that I've halfway finished a journey. I expect this essay to be most useful to people similarly-shaped-to-me, who are also undergoing that journey and could use some reassurance that there's an actual destination worth striving for.

Epistemic Status starts out "true story", and gets more (but not excessively) speculative with each section.

i. Gratitude

> "*Rationalists obviously don't *actually* take ideas seriously. Like, take the Gratitude Journal. This is the one peer-reviewed intervention that *actually increases your subjective well being*, and costs barely anything. And no one I know has even seriously tried it. Do literally *none* of these people care about their own happiness?*"
>
> "*Huh. Do *you* keep a gratitude journal?*"
>
> "*Lol. No, obviously.*"
>
> *- Some Guy at the Effective Altruism Summit of 2012*

Upon hearing the above, I decided to try gratitude journaling. It took me a couple years and a few approaches to get it working.

1. First, I **tried keeping a straightforward journal,** but it felt effortful and dumb.
2. I tried a thing where I ***wrote a poem* about the things I was grateful for,** but my mind kept going into "*constructing a poem*" mode instead of "*experience nice things mindfully*" mode.
3. I tried **just being mindful without writing anything down.** But I'd forget.
4. I tried **writing gratitude letters to people**, but it only occasionally felt right to do so. (This came *after* someone actually wrote *me* a handwritten gratitude letter, which felt amazing, but it felt a bit forced when I tried it myself.)
5. I tried **doing gratitude before I ate meals**, but I ate "real" meals inconsistently so it didn't take. (Upon reflection, maybe I should have fixed the "not eating real meals" thing?)

But then I stumbled upon something that worked. It was a *social* habit, which I worry is a bit fragile. I did it together with my girlfriend each night. On nights when one of us travelled, I'd often forget.

But this is the thing that worked. Each night, we share our Grumps and Gratitudes.

Grumps and Gratitudes goes like this:

1. We share anything we're annoyed or upset about. (We call this The Grump. Our rule is to not go *searching* for the Grump, simply to let it out if it's festering so that when we get to the Gratitude we actually appreciate it instead of feeling forced.)
2. We share three things that we're grateful for that day. On some bad days this is hard, but we should at least be able to return to old-standbys ("I'm breathing", "I have you with me"), and we always perform the action of at least *attempting* an effortful search.
3. Afterwards, pause to actually *feel* the Grates. Viscerally remember the thing and why it was nice. If we're straining to feel grateful and had to sort of reach into the bottom of the barrel to find something, we at least *try* to cultivate a mindset where we fully appreciate that thing.

Maybe the sun just glinted off your coffee cup nicely, and maybe that didn't stop the insurance company from screwing you over and your best friend from getting angry at you and your boss from firing you today.

But... in all seriousness... in a world whose laws of physics had no reason to make life even possible, a universe mostly full of empty darkness and no clear evidence of alien life out there, where the only intelligent life we know of sometimes likes to play chicken with nuclear arsenals...

...somehow some tiny proteins locked together ever so long ago and life evolved and consciousness evolved and somehow beauty evolved and... and here you are, a meatsack cobbled together by a blind watchmaker, and the sunlight is glinting off that coffee cup, and it's beautiful.

Over the years, I've gained an important related skill: *noticing* the opportunity to feel gratitude, and mindfully appreciating it.

I started writing this article because of a specific moment: I was sitting in my living room around noon. The sun suddenly filtered in through the window and, and on this particular day it somehow seemed achingly beautiful to me. I stared at it for 5 minutes, happy.

It seemed almost golden, in the Robert Frost sense. *Weirdly* golden.

It was like a sunset at noon.

(My coffee cup at 12:35pm. Photo does not capture the magic, you had to be there.)

And that might have been the entire essay here - a reminder to maybe cultivate gratitude (because it's, like, peer reviewed and *hopefully* hasn't failed to replicate), and to keep trying even if it doesn't seem to stick.

But I have a few more things on my mind, and I hope you'll indulge me.

ii. Tortoise Skills

Recently I read an article about a man living in India, near a desert sand bar. When he was 14 he decided that, every day, he would go there to plant a tree. Over time, those trees started producing seeds of their own. By taking root, they helped change the soil so that other kinds of plants and animals could live there.

Fifteen years later, the desert sandbar had become a forest as large as Central Park.

It's a cute story. It's a reminder that small, consistent efforts can add up to something meaningful. It also asks an interesting question:

Is whatever you're going to do for the next 15 years going to produce something at least as cool as a Central Park sized forest?

(This is not actually the forest in question, it's the image I could find easily that looked similar that was filed under creative commons. Credited to Your Mildura .)

A Three Percent Incline

A couple months ago, suddenly I noticed that... I had my shit together.

This was in marked contrast to 5 years ago when I decidedly *didn't* have my shit together:

- I struggled to stay focused at work for more than 2 hours at a time.
- I vaguely felt like I should exercise, but I didn't.
- I vaguely felt like I should be doing more productive things with my life, but I didn't.
- Most significantly, for the first three years of my involvement with the rationalsphere, I got *less happy, more stressed out, and seemed to get* worse *at thinking.* Valley of Bad Rationality indeed.

I absorbed the CFAR mantra of "try things" and "problems can in principle be factored into pieces, understood, and solved." So I dutifully looked over my problems, and attempted to factor and understand and fix them.

I tried things. Lots of things.

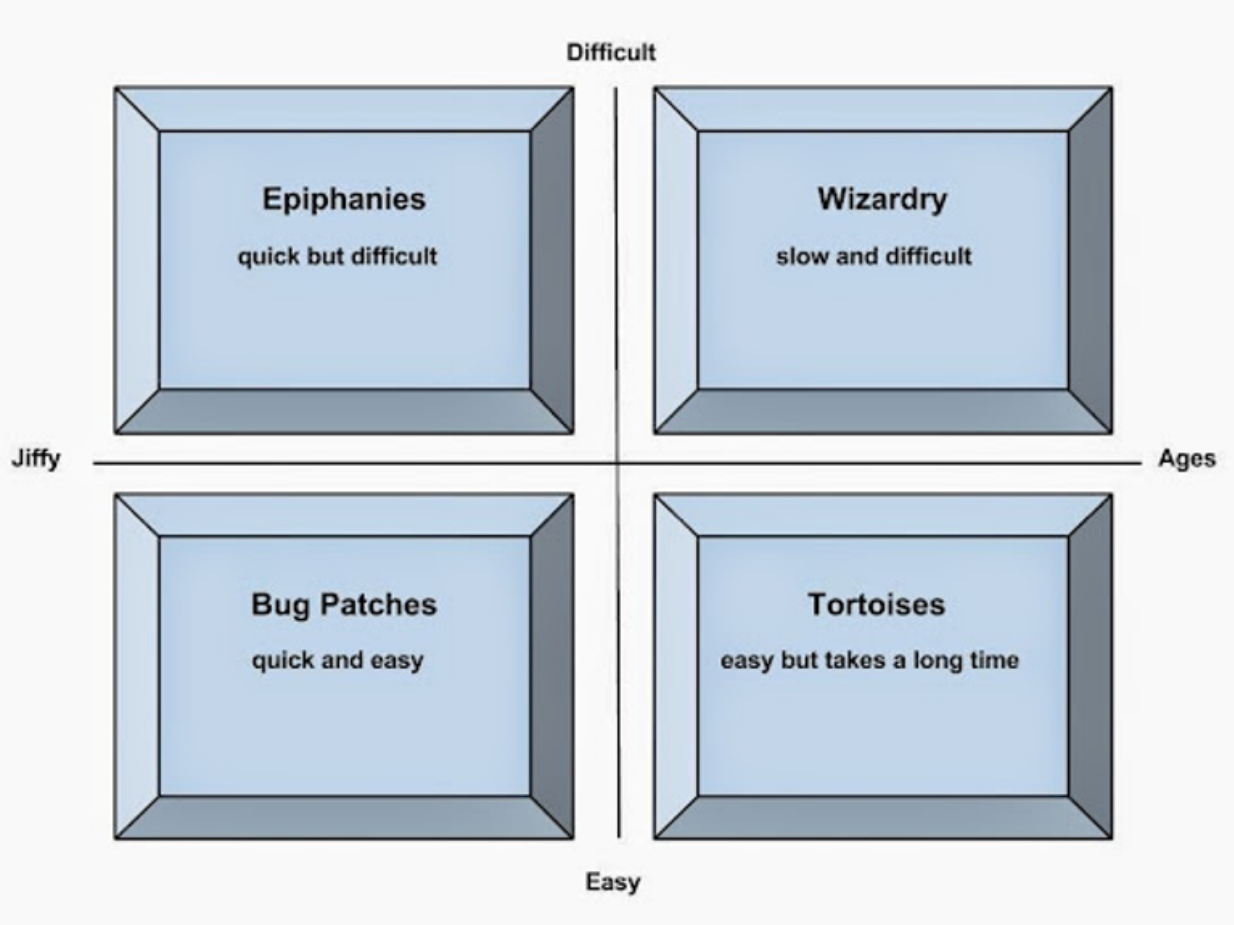
- I tried various systems and hacks to focus at work.
- I tried to practice mindfulness.
- I tried exercising - sometimes maintaining "1 pushup a day" microhabits. Sometimes major "work out at the gym" style things.
- I tried to understand my desires and bring conflicting goals into alignment so that I wasn't sabotaging myself.

My life did not especially change. Insofar as it did, it was because I undertook specific projects that I was excited about, which forced me to gain skills.

Years passed.

Somewhere in the middle of this, 2014, Brienne Yudkowsky wrote an essay about Tortoise Skills.

She divided skills into four quadrants, based on whether a skill was *fast* to learn, and how *hard* it was to learn.



LessWrong has (mostly) focused on *epiphanies* - concepts that might be difficult to get, but once you understand them you pretty much immediately understand them.

CFAR ends up focusing on epiphanies and *skills that can be taught in a single weekend,* because, well, they only have a single weekend to teach them. *Fully* gaining these skills takes a lot of practice, but in principle you can learn them in an hour.

There's some discussion about something you might call *Bayesian Wizardry* - a combination of deep understanding of probability, decision theory and 5-second reflexes. This seems very hard and takes a long time to see much benefit from.

But there seemed to also be an underrepresented "easy-but-time-consuming" cluster of skills, where the main obstacle was being *slow but steady.* Brienne went on to chronicle an exploration of deliberate habit acquisition, inspired by a similar project by Malcolm Ocean.

I read Brienne and Malcolm's works, as well as the book Superhuman by Habit, of which this passage was most helpful to me:

> Habits can only be thought of rationally when looked at from a perspective of years or decades. The benefit of a habit isn't the magnitude of each individual action you take, but the cumulative impact it will have on your life in the long term. It's through that lens that you must evaluate which habits to pick up, which to drop, and which are worth fighting for when the going gets tough.

> Just as it would be better to make 5% interest per year on your financial investments for the rest of your life than 50% interest for one year.... it's better to maintain a modest life-long habit than to start an extreme habit that can't be sustained for a single year.
>
> The practical implications of this are twofold.
>
> First, be conservative when sizing your new habits. Rather than say you will run every single day, agree to jog home from the train station every day instead of walk, and do one long run every week.
>
> Second, you should **be very scared to fail to execute a habit, even once.**
>
> By failing to execute, potentially you're not just losing a minor bit of progress, but rather threatening the cumulative benefits you've accrued by establishing a habit. This is a huge deal and should not be treated lightly. So make your habits relatively easy, but never miss doing them.
>
> **Absolutely *never* skip twice.**
>
> I was talking to a friend about a daily habit that I had. He asked me what I did when I missed a day. I told him about some of my strategies and how I tried to avoid missing a day. "What do you do when you miss two days?" he asked.
>
> "I don't miss two days," I replied.
>
> Missing two days of a habit is habit suicide. If missing one day reduces your chances of long-term success by a small amount like five percent, missing two days reduces it by forty percent or so.

"*Never miss 2 days*" was inspirational in a way that most other habit-advice hadn't been (though this may be specific to me). It had the "tough but fair coach is yelling at you" thing that some people find valuable, in a way that clearly had my long-term interests at heart.

So I started investing in habit-centric thinking. And it *still* wasn't super clear at first that anything good was really happening as a result...

...until suddenly, I looked back at my 5-years-ago-self...

...and noticed that I had my shit together.

It was like I'd been walking for 2 years, and it *felt* like I'd been walking on a flat, straight line. But in fact, that line had a 3% incline. And after a few years of walking I looked back and noticed I'd climbed to the top of a hill.

(Also as part of the physical exercise thing sometimes I climb literal hills.)

Some specific habits and skills I've acquired:

- I cultivate gratitude, floss, and do a few household chores every single day.
- I am able to focus at work for 4-6 hours instead of 2-4 (and semi-frequently get into the zone and do a full 8).
- Instead of "will I exercise at all today?", the question is more like "will I get around to doing 36 pushups today, or just 12?"
- I meditate for 5 minutes on most days.
- I have systems to ensure I get important things done, and a collection of habits that makes sure that important things end up in those systems.
- I'm *much more aware* of my internal mental states, and the mental states of people I interact with. I have a sense of what they mean, and what to do when I notice unhealthy patterns.
- Perhaps most importantly: the habit of *trying things* that seem like they might be helpful, and occasionally discovering something important, like improv, or like the website freedom.to.

On the macro level, I'm *more the sort of person who deliberately sets out to achieve things,* and follow through on them. And I'm able to do it while being generally happy, which didn't use to be the case. (This largely involves being comfortable not pushing myself, and guarding my slack.)

So if you've been trying things sporadically, and don't feel like you're moving anywhere, I think it's worth keeping in mind:

1. Are you aiming for *consistency* - making sure not to drop the ball on the habits you cultivate, however small?
2. If you've been trying things for a while, and it doesn't feel like you're making progress, it's worth periodically looking back and checking how far you've come.

Maybe you *haven't* been making progress (which is indeed a warning sign that something isn't working). But maybe you've just been walking at a steady, slight incline.

Have you been climbing a hill? If you were to *keep climbing*, and you imagine decades of future-yous climbing further at the same rate as you, how far would they go?

iii. Bayesian Wizardry

> "*What do you most often do instead of thinking? What do you imagine you could do instead?*"
>
> *- advice a friend of mine got on Facebook, when asking for important things to reflect on during a contemplative retreat.*

I could stop the essay *here* too. And it'd be a fairly coherent "*hey guys maybe consider cultivating habits and sticking with them even when it seems hard? You too could be grateful for life and also productive isn't that cool?*"

But there is more climbing to do. So here are some hills I'm *currently* working on, which I'm finally starting to grok the importance of. And because I've seen evidence of 3% inclines yielding real results, I'm more willing to lean into them, even if they seem like they'll take a while.

I've had a few specific mental buckets for "what useful stuff comes out of the rationalsphere," including:

Epistemic fixes that are practically useful in the shortish term *(i.e. noticing when you are 'arguing for a side' instead of actually trying to find the truth).*

Instrumental techniques, which mostly amount to 'the empirically valid parts of self-help' *(i.e. Trigger Action Plans).*

Deep Research and Bayesian Wizardry (i.e. high quality, in depth thinking that pushes the boundary of human knowledge forward while paying strategic attention to what things matter most, working with limited time and evidence).

Orientation Around Important Things (i.e. once someone has identified something like X-Risk as a crucial research area, people who *aren't* interested in specializing their lives around it can still help out with practical aspects, like getting a job as an office manager).

Importantly, it used to seem like Deep Research and Bayesian Wizardry was something *other people did.* I did not seem smart enough to contribute.

I'm still not sure *how much* it's possible for me to contribute - there's a power law of potential value, and I clearly wouldn't be in the top tiers even if I dedicated myself fully to it.

But, in the past year, there's been a zeitgeist initiated by Anna Salamon that *being good at thinking seems useful,* and if you could only carve out time to *actually think* (and to practice, improving at it over time) maybe you could actually generate something worthwhile.

So I tried.

Earlier this year I carved out 4 hours to actually think about X-Risk, and I output this blogpost on what to do about AI Safety if you seem like a moderately smart person with no special technical aptitudes.

It wasn't the most valuable thing in the world, but it's been cited a few times by people I respect, and I think it was probably the most valuable 4 hours I've spent to date.

Problems Worth Solving

I haven't actually carved out time to think in the same way since then - a giant block of time dedicated to a concrete problem. It may turn out that I used up the low-hanging fruit there, or that it requires a year's worth of conversations and shower-thoughts in order to build up to it.

But I look at people like Katja Grace - who just sit and actually look at what's going on with computer hardware, or come up with questions to ask actual AI researchers about what progress they expect. And it seems like there's a lot of things worth doing that *don't require you to have any weird magic*. You should just need to actually think about it, and then actually follow that thinking up with action.

I've also talked more with people who *do* seem to have something like weird magic, and I've gotten more of a sense that the magic has gears. It works for comprehensible reasons. I can see how the subskills build into larger skills. I can see the broad shape of how those skills combine into a cohesive source of cognitive power.

A few weeks ago, I was arguing with someone about the relative value of LessWrong (as a conversational locus of quality thinking) versus donating money to other causes. I can't remember their exact words, but a paraphrase:

> It's approximately as hard to have an impact by donating as by thinking - especially now that the effective altruism funding ecosystem has become more crowded. There are billions of dollars available - the hard part is knowing what to do with them. And often, when the answer is "use them to hire researchers to think about things", you're still passing the recursive buck.
>
> Someone has to think. And it's about as hard to get good at thinking as it is to get rich.

Meanwhile, some other conversations I've had with people in the EA, X-Risk and Rationality communities could be combined and summarized as:

> We have a lot of people showing up, saying "I want to help." And the problem is, the thing we most need help with is figuring out what to do. We need people with

> breadth and depth of understanding, who can look at the big picture and figure out what needs doing
>
> This applies just as much to "office manager" type positions as "theoretical researcher" types.

iv. Noticing Confusion

Brienne has a series of posts on Noticing Things, which is among the most useful, practical writings on *epistemic* rationality that I've read.

It notes:

> I suspect that the majority of good epistemic *practice* is best thought of as cognitive trigger-action plans.
>
> **[If I'm afraid of a proposition] → [then I'll visualize how the world would be and what I would actually do if the proposition were true.]**
>
> **[If everything seems to hang on a particular word] → [then I'll taboo that word and its synonyms.]**
>
> **[If I flinch away from a thought at the edge of peripheral awareness] → [then I'll focus my attention directly on that thought.]**

She later remarks:

> I was at first astonished by how often my pesky cognitive mistakes were solved by nothing but skillful use of attention. Now I sort of see what's going on, and it feels less odd.
>
> What happens to your bad habit of motivated stopping when you train consistent reflective attention to "motivated stopping"? The motivation dissolves under scrutiny...
>
> If you recognize something as a mistake, part of you probably has at least some idea of what to do instead. Indeed, *anything besides ignoring the mistake* is often a good thing to do instead. So merely noticing when you're going wrong can be over half the battle.

She goes on to chronicle her own practice at training the art of noticing.

This was helpful to me, and one particular thing I've been focusing lately is *noticing confusion.*

In the Sequences and Methods of Rationality, Eliezer treats "noticing confusion" like a sacred phrase of power, whispered in hushed tones. But for the first 5 or so years of my participation in the rationality community, I didn't find it that *useful*.

Confusion Is Near-Invisible

First of all, confusion (at least as I understand Eliezer to use the term) is *hard* to notice. The phenomenon here is when bits of evidence *don't add up*, and you get a subtle

sense of wrongness. But then instead of heeding that wrongness and making sense of it, you round the evidence to zero, or you round the situation to the nearest plausible cliché.

Some examples of confusion are simple: CFAR's epistemic habit checklist describes a person who thought they were supposed to get on a plane on Thursday. They got an email on Tuesday reminding them of their flight "tomorrow." This seemed odd, but their brain brushed it off as a weird anomaly that didn't matter.

In this case, noticing confusion is straightforwardly useful - you miss fewer flights.

Some instances are harder. A person is murdered. Circumstantial evidence points on one particular murderer. But there's a tiny note of discord. The evidence doesn't *quite* fit. A jury that's tired and wants to go home is looking for excuses to get the sentencing over with.

Sometimes it's harder still: you tell yourself a story about how consciousness works. It feels satisfactory. You have a brief flicker of awareness that your story doesn't explain consciousness well enough that you could *build it from scratch,* or discern when a given clump of carbon or silicon atoms would start being able to *listen* in a way that matters.

In this case, **it's not enough to *notice* confusion**. You have to follow it up with the **hard work of *resolving it.***

You may need to brainstorm ideas, validate hypotheses. To find the answer fastest and most accurately, you may need to not just "remember base rates", but to actually think about Bayesian probability as you explore those hypotheses with scant evidence to guide you.

Noticing confusion can be a tortoise skill, if you seek out opportunities to practice. But *doing something* with that confusion requires some wizardry.

(Incidentally: in at least one point earlier in this essay, if I told you you were given the opportunity to practice noticing confusion, could you identify where it was?)

v. The World Is Literally On Fire

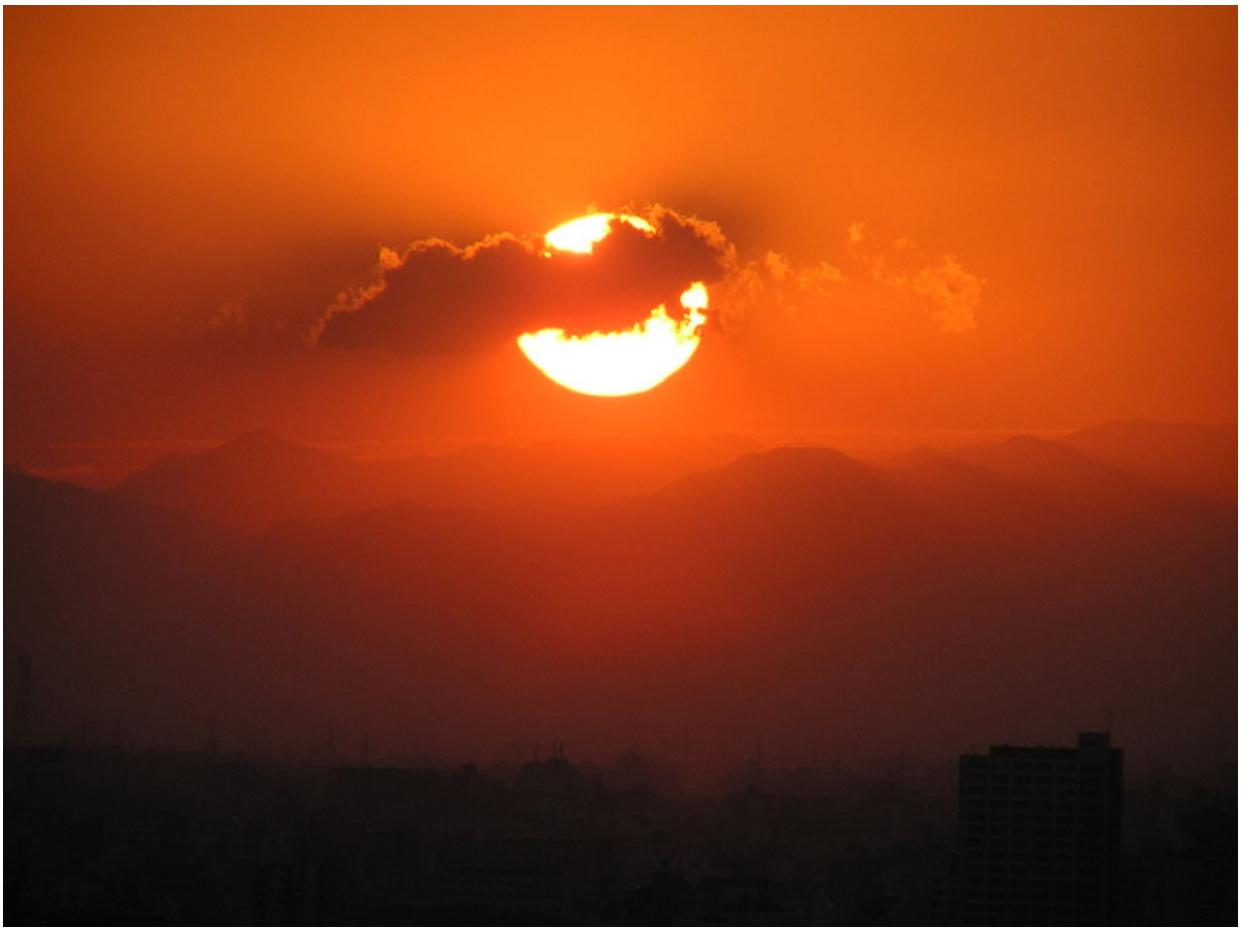
I've gotten pretty good at noticing when *I should* have been confused*,* after the fact.

A couple weeks ago, I was walking around my neighborhood. I smelled smoke.

I said to myself: "*huh, weird.*" An explanation immediately came to mind - someone was having a barbecue.

I do think this was the most likely explanation given my knowledge at the time. Nonetheless, it is *interesting* that a day later, when I learned that many nearby towns in California were literally on fire, and the entire world had a haze of smoke drifting through it... I thought back to that "*huh, weird*."

Something *had* felt out of place, and I *could* have noticed. I'd been living in Suburbia for a month or two and not noticed this smell, and while it *probably* was a barbecue, something about this *felt off.*



(When the world's on fire, the sun pretty unsubtly declares that things are not okay)

Brienne actually look this a step farther in a Facebook thread, paraphrased:

> "I notice that I'm confused about the California Wildfires. There are a *lot* of fires, all across the county. Far enough apart that they can't have spread organically. Are there often wildfires that spring up at the same time? Is this just coincidence? Do they have a common cause?"

Rather than stop at "notice confusion", she and people in the thread went on to discuss hypotheses. Strong winds were reported. Were they blowing the fires across the state? That still seemed wrong - the fires were skipping over large areas. Is it because California is in a drought? This explains why it's *possible* for lots of fires to abruptly start. But doesn't explain why they all started *today.*

The consensus eventually emerged that the fires had been caused by electrical sparks - the common cause was the strong winds, which caused powerlines to go down *in multiple locations.* And *then,* California being a dry tinderbox of fuel enabled the fires to catch.

I don't know if this is the true answer, but my own response, upon learning about the wildfires and seeing the map of where they were, had simply been, "huh." My curiosity stopped, and I didn't even *attempt* to generate hypotheses that adequately explained anything.

There are very few opportunities to practice noticing confusion.

When you notice yourself going "huh, weird" in response to a strange phenomenon... maybe that particular moment isn't that important. I certainly didn't change my actions due to understanding what caused the fires. But you *are* being given a scarce resource - the chance, in the wild, to *notice what noticing confusion feels like.*

Generating/evaluating hypotheses can be done in response to artificial puzzles and abstract scenarios, but the initial "*huh*" is hard to replicate, and I think it's important to train not just to notice the "huh" but to follow it up with the harder thought processes.

vi. ...also, Metaphorically On Fire

It so happened that this was the week that Eliezer published There Is No Fire Alarm for Artificial General Intelligence.

> In the classic experiment by Latane and Darley in 1968, eight groups of three students each were asked to fill out a questionnaire in a room that shortly after began filling up with smoke. Five out of the eight groups didn't react or report the smoke, *even as it became dense enough to make them start coughing*. Subsequent manipulations showed that a lone student will respond 75% of the time; while a student accompanied by two actors told to feign apathy will respond only 10% of the time.
>
> The fire alarm doesn't tell us with certainty that a fire is there. In fact, I can't recall one time in my life when, exiting a building on a fire alarm, there was an actual fire. Really, a fire alarm is *weaker* evidence of fire than smoke coming from under a door.
>
> But the fire alarm tells us that it's socially okay to react to the fire. It promises us with certainty that we won't be embarrassed if we now proceed to exit in an orderly fashion.

In typically Eliezer fashion, this would all be a metaphor for how there's *not* ever going to be a moment when it feels socially, professionally safe to be publicly worried about AGI.

Shortly afterwards, Alpha Go Zero was announced to the public.

For the past 6 years, I've been reading the arguments about AGI, and they've sounded plausible. But most of those arguments have involved a lot of metaphor and it seemed likely that a clever arguer could spin something similarly-convincing but false.

I did a lot of hand wringing, listening to Pat Modesto-like voices in my head. I eventually (about a year ago) decided the arguments were sound enough that I should move from the "think about the problem" to "actually take action" phase.

But it still didn't *really* seem like AGI was a real thing. I believed. I didn't alieve.

Alpha Go Zero changed that, for me. For the first time, the arguments were clear-cut. There was not just theory but concrete evidence that learning algorithms could improve quickly, that architecture could be simplified to yield improvement, that you could go from superhuman to super-super-human in a year.

Intellectually, I'd loosely believed, based on the vague authority of people who seemed smart, that maybe we might all be dead in 15 years.

And for the first time, seeing the gears laid bare, I felt the weight of alief that our civilization might be cut down in its prime.

...

(Incidentally, a few days later I was at a friends' house, and we smelled something vaguely like gasoline. Everyone said "huh, weird", and then turned back to their work. On this particular occasion I said "Guys! We JUST read about fire alarms and how people won't flee rooms with billowing smoke and CALIFORNIA IS LITERALLY ON FIRE RIGHT NOW. Can we look into this a bit and figure out what's going on?"

We then examined the room and brainstormed hypotheses and things. On this occasion we did not figure anything out and eventually the smell went away and we shrugged and went back to work. This was not the most symbolically useful anecdote I could have hoped for, but it's what I got.)

vii. Burning Out

People vary in what they care about, and how they naturally handle that caring. I make no remark on what people *should* care about.

But if you're shaped something like me, it may seem like the world is on fire at multiple levels. AI seems around 15% likely to kill everyone in 15 years. If it weren't, people around the world would still be dying for stupid preventable reasons, and people around the world would still be *living* but cut off from their potential.

Meanwhile, civilization seems disappointingly dysfunctional in ways that turn stupid, preventable reasons into confusing, intractable ones.

The metaphorical fires I notice range in order-of-magnitude-of-awfulness, but each seems sufficiently alarming that it completely breaks my grim-o-meter and renders it useless.

For three years, the rationality and effective altruism movements made me less happy, more stressed out, in ways that were clearly unsustainable and pointless.

The world is burning, but burning out doesn't help.

I don't have a principled take on how to integrate all of that. Some people have techniques that work for them. Me, I've just developed crude coping mechanisms of "stop feeling things when they seem overwhelming."

I do recommend that you guard your slack.

And if personal happiness is a thing you care about, I do recommend cultivating gratitude. Even when it turns out the reason your coffee cup was delightfully golden was that the world was burning.

Do what you think needs doing, but no reason not to be cheerful about it.

viii. Sunset at Noon

Earlier, I noted my coffee cup was beautiful. *Weirdly* beautiful. Like a sunset at noon.

That is essentially, verbatim, the series of thoughts that passed through my head, giving you approximately as much opportunity to pay attention as I had.

If you noticed that *sunsets are not supposed to happen at noon,* bonus points to you. If you stopped to hypothesize *why,* have some more. (I did neither).

Sometimes, apparently, the world is just literally on fire and the sky is covered in ash and the sun is an apocalyptic scareball of death and your coffee cup is pretty.

Sometimes you are lucky enough for this not to matter much, because you live safely a few hours' drive away, and your friends and the news and weather.com all let you know.

Sometimes, maybe you don't have time for friends to let you know. You're living an hour away from a wildfire that's spreading fast. And the difference between escaping alive and asphyxiating is having trained to *notice* and *act on* the small note of discord as the thoughts flicker by:

"Huh, weird."

(To the right: what my coffee cup normally looks like at noon)

Moloch's Toolbox (1/2)

Follow-up to: An Equilibrium of No Free Energy

There's a toolbox of reusable concepts for analyzing systems I would call "inadequate"—the causes of civilizational failure, *some* of which correspond to local opportunities to do better yourself. I shall, somewhat arbitrarily, sort these concepts into three larger categories:

1. Decisionmakers who are not beneficiaries;

2. Asymmetric information;

and above all,

3. Nash equilibria that aren't even the best Nash equilibrium, let alone Pareto-optimal.

In other words:

1. Cases where the decision lies in the hands of people who would gain little personally, or lose out personally, if they did what was necessary to help someone else;

2. Cases where decision-makers can't reliably learn the information they need to make decisions, even though someone else has that information; and

3. Systems that are broken in multiple places so that no one actor can make them better, even though, in principle, some magically *coordinated* action could move to a new stable state.

I will then play fast and loose with these concepts in order to fit the entire Taxonomy of Failure inside them.

For example, "irrationality in the form of cognitive biases" wouldn't *obviously* fit into any of these categories, but I'm going to shove it inside "asymmetric information" via a clever sleight-of-hand. Ready? Here goes:

If *nobody* can detect a cognitive bias in particular cases, then from our perspective we can't really call it a "civilizational inadequacy" or "failure to pluck a low-hanging fruit." We shouldn't even be able to see it ourselves. So, on the contrary, let's suppose that you and some other people can indeed detect a cognitive bias that's screwing up civilizational decisionmaking.

Then why don't you just walk up to the decision-maker and *tell* them about the bias? Because they wouldn't have any way of knowing to trust *you* rather than the other five hundred people trying to influence their decisions? Well, in that case, you're holding information that they can't learn from you! So that's an "asymmetric information problem," in much the same way that it's an asymmetric information problem when you're trying to sell a used car and *you* know it doesn't have any

mechanical problems, but you have no way of reliably conveying this knowledge to the buyer because for all they know you could be lying.

That argument is a bit silly, but so is the notion of trying to fit the whole Scroll of Woe into three supercategories. And if I named more than three supercategories, you wouldn't be able to remember them due to computational limitations (which aren't on the list anywhere, and I'm not going to add them).

i. For want of docosahexaenoic acids, a baby was lost

My discussion of modest epistemology in Chapter 1 might have given the impression that I think of modesty mostly as a certain set of high-level beliefs: beliefs about how best to combat cognitive bias, about how individual competencies stack up against group-level competencies, and so on. But I predict that many of this book's readers have high-level beliefs similar to those I outlined in Chapter 2, while employing a reasoning style that is really a special case of modest epistemology; and I think that this reasoning style is causing them substantial harm.

As reasoning styles, modest epistemology and inadequacy analysis depend on a mix of explicit principles and implicit mental habits. In inadequacy analysis, it's one thing to recognize in the abstract that we live in a world rife with systemic inefficiencies, and quite another to naturally *perceive* systems that way in daily life. So my goal here won't be to unkindly stick the label "inadequate" to a black box containing the world; it will be to say something about how the relevant systems actually operate.

For our central example, we'll be using the United States medical system, which is, so far as I know, the most broken system *that still works* ever recorded in human history. If you were reading about something in 19th-century France which was as broken as US healthcare, you wouldn't expect to find that it went on working when overloaded with a sufficiently vast amount of money. You would expect it to just not work at all.

In previous years, I would use the case of central-line infections as my go-to example of medical inadequacy. Central-line infections, in the US alone, killed 60,000 patients per year, and infected an additional 200,000 patients at an average treatment cost of $50,000/patient.

Central-line infections were also known to decrease by 50% or more if you enforced a five-item checklist that included items like "wash your hands before touching the line."

Robin Hanson has old *Overcoming Bias* blog posts on that untaken, low-hanging fruit. But I discovered while re-Googling in 2015 that wider adoption of hand-washing and similar precautions are now finally beginning to occur, after many years—with an associated 43% *nationwide* decrease in central-line infections. After *partial* adoption.[1]

So my new example is infants suffering liver damage, brain damage, and death in a way that's even easier to solve, by changing the lipid distribution of parenteral nutrition to match the proportions in breast milk.

Background: Some babies have digestion problems that require direct intravenous feeding. Long ago, somebody created a hospital formula for this intravenous feeding that matched the distribution of “fat,” “protein,” and “carbohydrate” in breast milk.

Just like “protein” comes in different amino acids, some of which the body can’t make on its own and some of which it can, what early doctors used to think of as “fat” actually breaks down into metabolically distinct elements like short-chain triglycerides, medium-chain triglycerides, saturated fat, and omega-6, omega-9, and the famous “omega-3.” “Omega-3” is actually several different lipids in its own right; vegetable oils with “omega-3” usually just contain alpha-linolenic acids, which can only be inefficiently converted to ecosapentaenoic acids, which are then even more inefficiently converted to docosahexaenoic acids, which are the actual key structural components in the body. This conversion pathway is rate-limited by a process that also converts omega-6, so too much omega-6 can prevent you from processing ALA into DHA even if you’re getting ALA.

So what happens if your infant nutrition was initially designed based on the concept of “fat” as a natural category, and all the “fat” in the mix comes from soybean oil?

From a popular book by Jaminet and Jaminet:

> Some babies are born with “short bowel syndrome” and need to be given parenteral nutrition, or nutrition delivered intravenously directly to the blood, until their digestive tracts grow and heal. Since 1961, parenteral nutrition has used soybean oil as its source of fat.[6] And for decades, babies on parenteral nutrition have suffered devastating liver and brain damage. The death rate on soybean oil is 30 percent by age four. […]
>
> In a clinical trial, of forty-two babies given fish oil [after they had already developed liver damage on soybean oil], three died and one required a liver transplant; of forty-nine given soybean oil, twelve died and six required a liver transplant.[8] The death-or-liver-transplant rate was reduced from 37 percent with soybean oil to 9 percent with fish oil.[2]

When Jaminet and Jaminet wrote the above, in 2012, there was a single hospital in the United States that could provide correctly formulated parenteral nutrition, namely the Boston Children’s Hospital; nowhere else. This formulation was illegal to sell across state lines.

A few years *after* the Boston Children’s Hospital developed their formula—keeping in mind the heap of dead babies continuing to pile up in the meanwhile—there developed a shortage of “certified lipids” (FDA-approved “fat” for adding to parenteral nutrition). For a year or two, the parenteral nutrition contained *no fat at all* which is *worse* and can kill *adults*.

You see, although there’s nothing special about the soybean oil in parenteral nutrition, there was only one US manufacturer approved to add it, and that manufacturer left the market, so…

As of 2015, the state of affairs was as follows: The FDA eventually solved the problem with the shortage of US-certified lipids, by… allowing US hospitals to import parenteral nutrition bags from Europe. And it only took them two years’ worth of dead patients to figure that out!

As of 2016, if your baby has short bowel syndrome, and has *already* ended up with liver damage, and either you or your doctor is lucky enough to know what's wrong and how to fix it, your doctor can apply for a special permit to use a non-FDA-approved substance for your child on an emergency basis. After this, you can buy Omegaven and hope that it cures your baby and that there isn't too much permanent damage and that it's not already too late.

This is an improvement over the prior situation, where the non-poisonous formulation was illegal to sell across state lines under any circumstances, but it's still not *good* by any stretch of the imagination.

Now imagine trying to explain to a visitor from a relatively well-functioning world just why it is that your civilization has killed a bunch of babies and subjected other babies to pointless brain damage.

"It's not that we're *evil,*" you say helplessly, "it's that… well, you see, it's not that anyone *wanted* to kill those babies, it's just the way the System ended up, somehow…"

ii. Asymmetric information and lemons problems

Three people have gathered in a blank white space:

- The VISITOR from a Better World;
- SIMPLICIO , who is attending a major university but hasn't taken undergraduate economics;
- CECIE, the Conventional Cynical Economist.

The Visitor speaks first.

VISITOR: So I've listened to you explain about babies suffering death and brain damage from parenteral nutrition built on soybean oil. I have several questions here, but I'll start with the most obvious one.

CECIE: Go ahead.

VISITOR: Why aren't there riots?

SIMPLICIO: The first thing you have to understand, Visitor, is that the folk in this world are hypocrites, cowards, psychopaths, and sheep.

I mean, *I* certainly care about the lives of newborn children. Hearing about their plight certainly makes *me* want to do something about it. When I see the problem continuing in spite of that, I can only conclude that other people *don't* feel the level of moral indignation that I feel when staring at a heap of dead babies.

CECIE: I don't think that hypothesis is needed, Simplicio. As a start, Visitor, you have to realize that the picture I've shown you is not widely known. Maybe 10% of the population, at most, is walking around with the prior belief that the FDA in general is killing people; our government runs on majority rule and the 10% can't unilaterally

defy it.[3] Maybe 0.1% of that 10% know that omega-3 ALA is converted into omega-3 DHA via a metabolic pathway that competes with omega-6. And then most of those aren't aware of what's happening to babies right now.

VISITOR: Pointing to that state of ignorance is hardly a sufficient explanation! If a theater is on fire and only one person knows it, they yell "Fire!" and then more people know it. People from my civilization would scream "Babies are dying over here!" and other people from my civilization would whip around their heads and look.

SIMPLICIO: Our world's cowards and sheep would hear that and think that it's (a) somebody else's problem and (b) all part of the plan.

CECIE: In our world, Visitor, we have an economic phenomenon sometimes called the lemons problem. Suppose you want to sell a used car, and I'm looking for a car to buy. From my perspective, I have to worry that your car might be a "lemon"—that it has a serious mechanical problem that doesn't appear every time you start the car, and is difficult or impossible to fix. Now, *you* know that your car isn't a lemon. But if I ask you, "Hey, is this car a lemon?" and you answer "No," I can't trust your answer, because you're incentivized to answer "No" either way. Hearing you say "No" isn't much Bayesian evidence. *Asymmetric information* conditions can persist even in cases where, like an honest seller meeting an honest buyer, both parties have strong incentives for accurate information to be conveyed.

A further problem is that if the fair value of a non-lemon car is $10,000, and the possibility that your car is a lemon causes me to only be willing to pay you $8,000, you might refuse to sell your car. So the honest sellers with reliable cars start to leave the market, which further shifts upward the probability that any given car for sale is a lemon, which makes me less willing to pay for a used car, which incentivizes more honest sellers to leave the market, and so on.

VISITOR: What does the lemons problem have to do with your world's inability to pass around information about dead babies?

CECIE: In our world, there are a lot of people screaming, "Pay attention to this thing I'm indignant about over here!" In fact, there are enough people screaming that there's an inexploitable market in indignation. The dead-babies problem can't compete in that market; there's no free energy left for it to eat, and it doesn't have an optimal indignation profile. There's no single individual villain. The business about competing omega-3 and omega-6 metabolic pathways is something that only a fraction of people would understand on a visceral level; and even if those people posted it to their Facebook walls, most of their readers wouldn't understand and repost, so the dead-babies problem has relatively little virality. Being indignant about this particular thing doesn't signal your moral superiority to anyone else in particular, so it's not viscerally enjoyable to engage in the indignation. As for adding a further scream, "But wait, this matter *really is* important!", that's the part subject to the lemons problem. Even people who honestly know about a fixable case of dead babies can't emit a *trustworthy* request for attention.

SIMPLICIO: You're saying that people won't listen even if I sound *really* indignant about this? That's an outrage!

CECIE: By this point in our civilization's development, many honest buyers and sellers have left the indignation market entirely; and what's left behind is not, on average, good.

VISITOR: Your reply contains so many surprising postulates of weird civilizational dysfunction, I hardly know what to ask about next. So instead I'll try to explain how my world works, and you can explain to me why your world doesn't work that way.

CECIE: Sounds reasonable.

iii. Academic incentives and beneficiaries

VISITOR: To start with, in my world, we have these people called "scientists" who verify claims experimentally, and other people trust the "scientists." So if our "scientists" say that a certain formula seems to be killing babies, this would provoke general indignation without every single listener needing to study docohexa-whatever acids.

SIMPLICIO: Alas, our so-called scientists are just pawns of the same medical-industrial complex that profits from killing babies.

CECIE: I'm afraid, Visitor, that although there are strong prior reasons to expect too much omega-6 and no omega-3 to be very bad for an infant baby, and there are now a few dozen small-scale studies which seem to match that prediction, this matter hasn't had the massive study that would begin to produce confident scientific agreement—

VISITOR: You'd better not be pointing to *that* as an exogenous fact that explains your civilization's problem! See, on my planet, if somebody points to *strong prior suspicion* combined with *confirming pilot studies* saying that something is killing innocent babies and is fixable, and the pilot studies are not considered sufficient evidence to settle the issue, our people would *do more studies* and wouldn't just go on blindly feeding the babies poison in the meantime. Our scientists would all agree on *that*!

CECIE: But people loudly agreeing on something, by itself, accomplishes nothing. It's all well and good for everyone to agree in principle that larger studies ought to be done; but in your world, who actually does the big study, and why do they do it?

VISITOR: Two subclasses within the profession of "scientist" are *suggesters,* whose piloting studies provide the initial suspicions of effects, and *replicators* whose job it is to confirm the result and nail things down solidly—the exact effect size and so on. When an important suggestive result arises, two replicators step forward to confirm it and nail down the exact conditions for producing it, being forbidden upon their honor to communicate with each other until they submit their findings. If both replicators agree on the particulars, that completes the discovery. The three funding bodies that sustained the suggester and the dual replicators would receive the three places of honor in the announcement. Do I need to explain how part of the function of any civilized society is to appropriately reward those who contribute to the public good?

CECIE: Well, that's not how things work on Earth. Our world gives almost all the public credit and fame to the *discoverer*, as the initial suggester is called among us. Our scientists often *say* that replication is important, but our most prestigious journals won't publish mere replications; nor do the history books remember them. The outcome is a lot of small studies that have just enough subjects to obtain "statistically significant" results—

VISITOR: … What? Probability is quantitative, not qualitative. There's no such thing as a "significant" or "insignificant" likelihood ratio—

CECIE: *Anyway*, while it might be good if larger studies were done, *the decisionmaker is not the beneficiary*—the people who did the extra work of a larger study, and funded the extra work of a larger study, would not receive fame and fortune thereby.

VISITOR: I must be missing something basic here. You do have multiple studies, right? When you have multiple bodies of data, you can multiply the likelihood functions from the studies' respective data to the hypotheses to obtain the meaning of the *combined* evidence—the likelihood function from *all* the data to the hypotheses.[4]

CECIE: I'm afraid you can't do that on Earth.

VISITOR: … Of course you can. It's a *mathematical theorem*. You can't possibly tell me *that* differs between our universes!

Yes, there are pitfalls for the especially careless. Sometimes studies end up being conducted under different circumstances, with the result that the naively computed likelihood functions don't have uniform relations to the hypotheses under consideration. In that case, blindly multiplying will give you a likelihood function that's nearly zero everywhere. But, I mean, if you just look at all the likelihood functions, it's pretty obvious when some of them are pointing in different directions and then you can *investigate that divergence*.

Either it makes sense to multiply all the likelihood functions and get out one massive evidential pointer, or else you *don't* get a sensible result when you multiply them and then you know something's wrong with your methods—

CECIE: I'm afraid our scientific community doesn't run on your world's statistical methods. You see, during the first half of the twentieth century, it became conventional to measure something called "*p*-values" which imposed a qualitative distinction between "successful" and "unsuccessful" experiments—

VISITOR: *That is still not an explanation.* Why not *change* the way you do things?

CECIE: Because somebody who tried using unconventional statistical methods, even if they were better statistical methods, wouldn't be able to publish their papers in the most prestigious journals. And then they wouldn't get hired. It's similar to the way that the most prestigious journals don't publish mere replications, only discoveries, so people focus on making discoveries instead of replications.

VISITOR: Why would anyone *pay attention* to journals like that?

CECIE: Because university hiring departments care a lot about whether you've published in prestigious journals.

VISITOR: No, I mean… how did these journals end up prestigious in the first place? *Why* do university hiring departments pay attention to them?

SIMPLICIO: Why *would* university hiring departments care about real science? Shouldn't it be you who has to explain why some lifeless cog of the military-industrial complex would care about anything except grant money?

CECIE: Okay… you're digging pretty deep here. I think I need to back up and try to explain things on a more basic level.

VISITOR: Indeed, I think you should. So far, every time I've asked you why someone is acting insane, you've claimed that it's secretly a sane response to someone else acting insane. Where does this process bottom out?

iv. Two-factor markets and signaling equilibria

CECIE: Let me try to identify a first step on which insanity can emerge from non-insanity. Universities pay attention to prestigious journals because of a *signaling equilibrium,* which, in our taxonomy, is a kind of bad Nash equilibrium that no single actor can defy unilaterally.

In your terms, it involves a sticky, stable equilibrium of *everyone* acting insane in a way that's secretly a sane response to everyone else acting insane.

VISITOR: Go on.

CECIE: First, let me explain the idea of what Eliezer has nicknamed a "two-factor market." Two-factor markets are a conceptually simpler case that will help us later understand signaling equilibria.

In our world there's a crude site for classified ads, called Craigslist. Craigslist doesn't contain any way of rating users, the way that eBay lets buyers and sellers rate each other, or that Airbnb lets renters and landlords rate each other.

Suppose you wanted to set up a version of Craigslist that let people rate each other. Would you be able to compete with Craigslist?

The answer is that even if this innovation is in fact a good one, competing with Craigslist would be far more difficult than it sounds, because Craigslist is sustained by a two-factor market. The sellers go where there are the most buyers; the buyers go where they expect to find sellers. When you launch your new site, no buyers will want to go there because there are no sellers, and no sellers will want to go there because there are no buyers. Craigslist initially broke into this market by targeting San Francisco particularly, and spending marketing effort to assemble the San Francisco buyers and sellers into the same place. But that would be harder to do for a later startup, because now the people it's targeting are already using Craigslist.

SIMPLICIO: Those sheep! Just mindlessly doing whatever their incentives tell them to!

CECIE: We can imagine that there's a better technology than Craigslist, called Danslist, such that everyone using Craigslist would be better off if they all switched to Danslist simultaneously. But if just one buyer or just one seller is the first to go to Danslist, they find an empty parking lot. In conventional cynical economics, we'd say that this is a *coordination problem—*

SIMPLICIO: A coordination problem? What do you mean by that?

CECIE: Backing up a bit: A “Nash equilibrium” is what happens when everyone makes their best move, given that all the other players are making their best moves from that Nash equilibrium—everyone goes to Craigslist, because that’s their individually best move *given* that everyone else is going to Craigslist. A “Pareto optimum” is any situation where it’s impossible to make every actor better off simultaneously, like “Cooperate/Cooperate” in the Prisoner’s Dilemma—there’s no alternative outcome to Cooperate/Cooperate that makes *both* agents better off. The Prisoner’s Dilemma is a coordination problem because the sole Nash equilibrium of Defect/Defect isn’t Pareto-optimal; there’s an outcome, Cooperate/Cooperate, that both players prefer, but aren’t reaching.

SIMPLICIO: How stupid of them!

CECIE: No, it’s… ah, never mind. Anyway, the *frustrating* parts of civilization are the times when you’re stuck in a Nash equilibrium that’s Pareto-inferior *to other Nash equilibria*. I mean, it’s not surprising that humans have trouble getting to non-Nash optima like “both sides cooperate in the Prisoner’s Dilemma without any other means of enforcement or verification.” What makes an equilibrium *inadequate,* a fruit that seems to hang tantalizingly low and yet somehow our civilization isn’t plucking, is when there’s a better *stable* state and we haven’t reached it.

VISITOR: Indeed. Moving from bad equilibria to better equilibria is the whole point of having a civilization in the first place.

CECIE: Being stuck in an inferior Nash equilibrium is how I’d describe the frustrating aspect of the two-factor market of buyers and sellers that can’t switch from Craigslist to Danslist. The scenario where everyone is using Danslist *would* be a stable Nash equilibrium, and a *better* Nash equilibrium. We just can’t get there from here. There’s no one actor who is behaving foolishly; all the individuals are responding strategically to their incentives. It’s only the larger system that behaves “foolishly.” I’m not aware of a standard term for this situation, so I’ll call it an “inferior equilibrium.”

SIMPLICIO: Why do you care what academics call it? Why not just use the *best* phrase?

CECIE: The terminology “inferior equilibrium” would be fine if everyone else were already using that terminology. Mostly I want to use the same phrase that everyone else uses, even if it’s not the best phrase.

SIMPLICIO: Regardless, I’m not seeing what the grand obstacle is to people solving these problems by, you know, *coordinating*. If people would just act in unity, so much could be done!

I feel like you’re placing too much blame on system-level issues, Cecie, when the simpler hypothesis is just that the people *in* the system are terrible: bad at thinking, bad at caring, bad at coordinating. You claim to be a “cynic,” but your whole world-view sounds rose-tinted to me.

VISITOR: Even in my world, Simplicio, coordination isn’t as simple as everyone jumping simultaneously every time one person shouts “Jump!” For coordinated action to be successful, you need to trust the institution that says what the action should be, and a *majority* of people have to trust that institution, and they have to *know* that other people trust the institution, so that everyone *expects* the coordinated action to occur at the critical time, so that it makes sense for them to act too.

That's why we have policy prediction markets and… there doesn't seem to be a word in your language for the *timed-collective-action-threshold-conditional-commitment…* hold on, this cultural translator isn't making any sense. "Kickstarter"? You have the key concept, but you use it mainly for making video games?

CECIE: I'll now introduce the concept of a *signaling equilibrium*.

To paraphrase a commenter on *Slate Star Codex*: suppose that there's a magical tower that only people with IQs of at least 100 and some amount of conscientiousness can enter, and this magical tower slices four years off your lifespan. The natural next thing that happens is that employers start to prefer prospective employees who have proved they can enter the tower, and employers offer these employees higher salaries, or even make entering the tower a condition of being employed at all.[5]

VISITOR: Hold on. There *must* be less expensive ways of testing intelligence and conscientiousness than sacrificing four years of your lifespan to a magical tower.

CECIE: Let's not go into that right now. For now, just take as an exogenous fact that employers can't get all of the information they want by other channels.

VISITOR: But—

CECIE: Anyway: the natural next thing that happens is that employers start to demand that prospective employees show a certificate saying that they've been inside the tower. This makes *everyone* want to go to the tower, which enables somebody to set up a fence around the tower and charge hundreds of thousands of dollars to let people in.[6]

VISITOR: But—

CECIE: Now, fortunately, after Tower One is established and has been running for a while, somebody tries to set up a competing magical tower, Tower Two, that also drains four years of life but charges less money to enter.

VISITOR: … You're *solving the wrong problem.*

CECIE: Unfortunately, there's a subtle way in which this competing Tower Two is hampered by the same kind of lock-in that prevents a jump from Craigslist to Danslist. Initially, all of the smartest people headed to Tower One. Since Tower One had limited room, it started discriminating further among its entrants, only taking the ones that have IQs above the minimum, or who are good at athletics or have rich parents or something. So when Tower Two comes along, the employers still *prefer* employees from Tower One, which has a more famous reputation. So the smartest people still prefer to apply to Tower One, even though it costs more money. This stabilizes Tower One's reputation as being the place where the smartest people go.

In other words, the signaling equilibrium is a two-factor market in which the stable point, Tower One, is cemented in place by the individually best choices of two different parts of the system. Employers prefer Tower One because it's where the smartest people go. Smart employees prefer Tower One because employers will pay them more for going there. If you try dissenting from the system unilaterally, without everyone switching at the same time, then as an employer you end up hiring the less-qualified people from Tower Two, or as an employee, you end up with lower salary offers after you go to Tower Two. So the system is stable as a matter of individual incentives, and stays in place. If you try to set up a cheaper alternative to the whole Tower system,

the *default* thing that happens to you is that people who couldn't handle the Towers try to go through your new system, and it acquires a reputation for non-prestigious weirdness and incompetence.

VISITOR: This all just seems so weird and complicated. I'm skeptical that this scenario with the magical towers could happen in real life.

SIMPLICIO: I agree that trying to build a cheaper Tower Two is solving the wrong problem. The interior of Tower One boasts some truly exquisite architecture and decor. It just makes sense that *someone* should pay a lot to allow people entry to Tower One. What we really need is for the government to subsidize the entry fees on Tower One, so that more people can fit inside.

CECIE: Consider a simpler example: Velcro is a system for fastening shoes that is, for at least some people and circumstances, better than shoelaces. It's easier to adjust three separate Velcro straps then it is to keep your shoelaces perfectly adjusted at all loops, it's faster to do and undo, et cetera, and not everyone is running at high speeds that call for perfectly adjusted running shoes. But when Velcro was introduced, the earliest people to adopt Velcro were those who had the most trouble tying their shoelaces—very young children and the elderly. So Velcro became associated with kids and old people, and thus unforgivably *unfashionable*, regardless of whether it would have been better than shoelaces in some adult applications as well.

VISITOR: I take it you didn't have the stern and upright leaders, what we call the Serious People, who could set an example by donning Velcro shoes themselves?

SIMPLICIO & CECIE: (*in unison*) No.

VISITOR: I see.

CECIE: Now consider the system of scientific journals that we were originally talking about. Some journals are prestigious. So university hiring committees pay the most attention to publications in that journal. So people with the best, most interesting-looking publications try to send them to that journal. So if a university hiring committee paid an equal amount of attention to publications in lower-prestige journals, they'd end up granting tenure to less prestigious people. Thus, the whole system is a stable equilibrium that nobody can unilaterally defy except at cost to themselves.

VISITOR: I'm still skeptical. Doesn't your parable of the magical tower suggest that, if that's actually true, somebody ought to rope off the journals too and charge insane amounts of money?

CECIE: Yes, and that's exactly what happened. Elsevier and a few other profiteers grabbed the most prestigious journals and started jacking up the access costs. They contributed almost nothing—even the peer review and editing was done by unpaid volunteers. Elsevier just charged more and more money and sat back. This is standardly called *rent-seeking*. In a few cases, the scientists were able to kickstart a coordinated move where the entire editing board would resign, start a new journal, and everybody in the field would submit to the new journal instead. But since our scientists don't have recognized kickstarting customs, or any software support for them, it isn't easy to pull that off. Most of the big-name journals that Elsevier has captured are still big names, still getting prestigious submissions, and still capturing big-money rents.

VISITOR: Well, I guess I understand why my cultural translator keeps putting air quotes around Earth's version of "science." The whole idea of science, as I understand the concept, is that everything has to be in the open for anyone to verify. Science is the part of humanity's knowledge that everyone can potentially learn about and reproduce themselves. You can't *charge money* in order for people to read your experimental results, or you lose the "everyone can access and verify your claims" property that distinguishes science from other kinds of information.

CECIE: Oh, rest assured that scientists aren't seeing any of this money. It all goes to the third-party journal owners.

SIMPLICIO: And this isn't just scientists being stupid?

CECIE: No stupider than you are for going to college. It's hard to beat *signaling equilibria*—because they're "multi-factor markets"—which are special cases of *coordination problems* that create "inferior Nash equilibria"—which are so stuck in place that market controllers can *seek rent* on the value generated by captive participants.

SIMPLICIO: Weren't we talking about dead babies at some point?

CECIE: Yes, we were. I was explaining how our system allocated too much credit to discoverers and not enough credit to replicators, and the only socially acceptable statistics couldn't aggregate small-scale trials in a way regarded as reliable. The Visitor asked me why the system was like that. I pointed to journals that published a particular kind of paper. The Visitor asked me why anyone paid attention to those journals in the first place. I explained about signaling equilibria, and that's where we are now.

VISITOR: I can't say that I feel enlightened at the end of walking through all that. There must be *particular* scientists on the editorial boards who choose not to demand replications and who forbid multiplying likelihood ratios. Why are those particular scientists doing the non-sensible thing?

CECIE: Because people in the general field wouldn't cite nonstandard papers, so if the editors demanded nonstandard papers, the journal's publication factor would decrease.

VISITOR: Why don't the journal editors start by demanding that paper submitters *cite* dual replications as well as initial suggestions?

CECIE: Because that would be a weird unconventional demand, which might lead people with high-prestige results to submit those results to other journals instead. Fundamentally, you're asking why scientists on Earth don't adopt certain new customs that you think would be for the good of everyone. And the answer is that there's this big, multi-factor system that nobody can dissent from unilaterally, and that people have a *lot* of trouble coordinating to change. That's true even when there are forces like Elsevier that are being *blatant* about ripping everyone off. Implementing your proposed cultural shift to "suggesters" and "replicators," or using likelihood functions, would be significantly *harder* than everyone just simultaneously ceasing to deal with Elsevier, since the case for it would be less obvious and would provoke more disagreement. All that we can manage is to make incremental shifts toward funding more replication and asking more for study preregistration.

To sum up, academic science is embedded in a big enough system with enough separate decisionmakers creating incentives for other decisionmakers that it almost always takes the path of least resistance. The system isn't in the *best* Nash equilibrium because nobody has the power to look over the system and choose *good* Nash equilibria. It's just in *a* Nash equilibrium that it wandered into, which includes statistical methods that were invented in the first half of the 20th century and editors not demanding that people cite replications.

VISITOR: I see. And that's why nobody in your world has multiplied the likelihood functions, or done a large-enough single study, or otherwise done *whatever it would take* to convince whoever needs to be convinced about the effects of feeding infants soybean oil.

CECIE: It's one of the reasons. A large study would also be very *expensive* because of extreme paperwork requirements, generated by other systemic failures I haven't gotten around to talking about yet—[7]

VISITOR: How does anything get done *ever*, in your world?

CECIE: —and when it comes to funding or carrying out that bigger study, *the decisionmaker would not significantly benefit* under the current system, which is held in place by *coordination problems*. And that's why people who already have a background grasp of lipid metabolic pathways have *asymmetric information* about what is worth becoming indignant about.

v. Total market failures

VISITOR: Even granting the things you've said already, I don't feel like I've been told enough to understand why your society is killing babies.

CECIE: Well, *no*. Not yet. The lack of incentive to do a large-scale convincing study is only *one* thing that went wrong inside *one* part of the system. There's a lot *more* broken than just that—which is why effective altruists shouldn't be running out and trying to fund a big replication study for Omegaven, because that by itself wouldn't fix things.

VISITOR: Okay, suppose there *had* been a large enough study to satisfy your world's take on "scientists." What *else* would likely go wrong after that?

CECIE: Several things. For example, doctors wouldn't necessarily be aware of the experimental results.

VISITOR: Hold on, I think my cultural translator is broken. You used that word "doctor" and my translator spit out a long sequence of words for Examiner plus Diagnostician plus Treatment Planner plus Surgeon plus Outcome Evaluator plus Student Trainer plus Business Manager. Maybe it's stuck and spitting out the names of all the professions associated with medicine.

CECIE: So, in your world, if there is a dual replication of results on Omegaven versus soybean oil, how does that end up changing the actual patient treatments?

VISITOR: By informing the Treatment Planners who specialize in infant ailments that required parenteral nutrition, of course. The discovery would appear inside the "parenteral nutrition" pages in the Earthweb and show up in the feeds of everyone subscribed to that page. The statistics would appear inside the Treatment Planner's decision-support software. And if all of those broke for some reason, every Treatment Planner for infant ailments that required parenteral nutrition would just use chatrooms. And anyone who ignored the chatrooms would have worse patient outcome ratings, and would lose status relative to Treatment Planners who were more attentive.

CECIE: It sounds like "Treatment Planners" in your world are much more specialized than doctors in this world. I suppose they're also selected specifically for talent at… cost-benefit analysis and decision theory, or something along those lines? And then they focus their learning on particular diseases for which they are Treatment Planners? And somebody else tracks their outcomes?

VISITOR: Of course. I'm… almost afraid to ask, but how do they do it in your world?

CECIE: Your translator wasn't broken. In our world, "doctors" are supposed to examine patients for symptoms, diagnose especially complicated or obscure ailments using their encyclopedic knowledge and their keen grasp of Bayesian inference, plan the patient's treatment by weighing the costs and benefits of the latest treatments, execute the treatments using their keen dexterity and reliable stamina, evaluate for themselves how well that went, train students to do it too, and in many cases, *also* oversee the small business that bills the patients and markets itself. So "doctors" have to be selected for all of those talents simultaneously, and then split their training, experience, and attention between them.

VISITOR: *Why* in the name of—

CECIE: Oh, and before they go to medical school, we usually send them off to get a four-year degree in philosophy first or something, just because.

I don't know if there's a standard name for this phenomenon, but we can call it "failure of professional specialization." It also appears when, for example, a lawyer has to learn calculus in order to graduate college, even though their job doesn't require any calculus.

VISITOR: Why. Why. Why why why—

CECIE: I'm not sure. I suspect the origin has something to do with status—like, a high-status person can do all things at once, so it's insulting and lowers status to suggest that an esteemed and respectable Doctor should only practice one surgical operation and get very good at it. And once you yourself have spent twelve years being trained under the current system, you won't be happy about the proposal to replace it with two years of much more specialized training. Once you've been through a painful initiation ritual and rationalized its necessity, you'll hate to see anyone else going through a less painful one. Not to mention that you won't be happy about the competition against your own human capital, by a cheaper and better form of human capital—and after the sunk cost in pain and time that you endured to build human capital under the old system…

VISITOR: Do they not have markets on your planet? Because on my planet, when you manufacture your product in a crazy, elaborate, expensive way that produces an

inferior product, someone else will come along and rationalize the process and take away your customers.

CECIE: We have markets, but there's this unfortunate thing called "regulatory capture," of which one kind is "occupational licensing."

As an example, it used to be that chairs were carefully hand-crafted one at the time by carpenters who had to undergo a lengthy apprenticeship, and indeed, they didn't like it when factories came along staffed by people who specialized in just carving a single kind of arm. But the factory-made chairs were vastly cheaper and most of the people who insisted on sticking to handcrafts soon went out of business.

Now imagine: What if the chair-makers had been extremely respectable—had already possessed very high status? What if their profession had an element of danger? What if they'd managed to frighten everyone about the dangers of improperly made chairs that might dump people on the ground and snap their necks?

VISITOR: Okay, yes, we used to have Serious People who would go around and certify the making of some medicines where somebody might be tempted to cheat and use inferior ingredients. But that was before *computers* and *outcome statistics* and *online ratings.*

CECIE: And on our planet, Uber and Lyft are currently fighting it out with taxi companies and their pet regulators after exactly that development. But suppose the whole system was set up before the existence of online ratings. Then the carpenters might have managed to introduce occupational licensing on who could be a carpenter. So if you tried to set up a factory, your factory workers would have needed to go through the traditional carpentry apprenticeship that covered every part of every kind of furniture, before they were legally allowed to come to your factory and specialize in carving just one kind of chair-arm. And then your factory would also need a ton of permits to sell its furniture, and would need to inveigle orders from a handful of resellers who were licensed to buy and resell furniture at a fixed margin. That small, insular group of resellers might not benefit *literally personally*—in their own personal salary—from buying from your cheaper factory system. And so it would go.

VISITOR: But why would the legislators go along with that?

CECIE: Because the carpenters would have a big, concentrated incentive to figure out how to make legislators do it—maybe by hiring very persuasive people, or by subtle bribery, or by not-so-subtle bribery.

Insofar as occupational licensing works to the benefit of professionals at the expense of consumers, occupational licensing represents a kind of regulatory capture, which happens when a few regulatees have a much more concentrated incentive to affect the regulation process. Regulatory capture in turn is a kind of commons problem, since every citizen shares the benefits of non-captured regulation, but no individual citizen has a sufficient incentive to unilaterally spend their life attending to that particular regulatory problem. So occupational licensing is regulatory capture is a commons problem is a coordination problem.

VISITOR: Then… the upshot is that it's impossible for your country to *test* a functional hospital design *in the first place?* The reformers can't win the competition because they're not legally allowed to try?

CECIE: But of course. Though in this case, if you did manage to set up a test hospital working along more reasonable lines, you still wouldn't be able to advertise your better results relative to any other hospitals. With just a few isolated exceptions, all of the other hospitals on Earth don't publish patient outcome statistics in the first place.

VISITOR: … But… then—*what are they even selling?*

SIMPLICIO: Hold on. If you reward the doctors with the highest patient survival rates, won't they just reject all the patients with poor prognoses?

VISITOR: Obviously you don't evaluate raw survival rates. You have Diagnosticians who estimate prognosis categories and are rated on their predictive accuracy, and Treatment Planners and Surgeons who are rated on their *relative* outcomes, and you have the outcomes evaluated by a third party, and—

CECIE: In our world, there's no separation of powers where one person assigns patients a prognosis category and has their prediction record tracked, and another person does their best to treat them and has their treatment record tracked. So hospitals don't publish any performance statistics, and patients choose the hospital closest to their house that takes their workplace's insurance, and nobody has any financial incentive to decrease the number of patient deaths from sloppy surgeons or central line infections. When anesthesiologists in particular did happen to start tracking patient outcomes, they adopted some simple monitoring standards and subsequently decreased their fatality rates by a factor of *one hundred*.[8] But that's just anesthesiologists, not, say, cardiac surgeons.

With cardiac surgeons, a group of researchers recently figured out how to detect when the most senior cardiac surgeons were at conferences, and found that the death rates went down while the most senior cardiac surgeons were away.[9] But our scientists have to use special tricks if they want to find out any facts like that.

VISITOR: Do your *patients* not care if they live or die?

CECIE: Robin Hanson has a further thesis about how what people really want from medicine is reassurance rather than statistics. But I'm not sure that hypothesis is necessary to explain this particular aspect of the problem. If no hospital offers statistics, then you have no baseline to compare to if one hospital *does* start offering statistics. You'd just be looking at an alarming-looking percentage for how many patients die, with no idea of whether that's a better percentage or a worse percentage. Terrible marketing! Especially compared to that other hospital across town that just smiles at you reassuringly.

No hospital would benefit from being the *first* to publish statistics, so none of them do.

VISITOR: Your world has literally zero market demand for empirical evidence?

CECIE: Not zero, no. But since publishing scary numbers would be bad marketing for *most* patients, and hospitals are heavily regional, they all go by the majority preference to not hear about the statistics.

VISITOR: I confess I'm having some trouble grasping the concept of a market consisting of opaque boxes allegedly containing goods, in which nobody publishes what is inside the boxes.

CECIE: Hospitals don't publish prices either, in most cases.

VISITOR: …

CECIE: Yeah, it's pretty bad even by Earth standards.

VISITOR: You literally don't *have* a healthcare market. Nobody knows what outcomes are being sold. Nobody knows what the prices are.

CECIE: I guess we could call that Total Market Failure? As in, things have gone so wrong that there's literally no supply-demand matching or price-equilibrating mechanism remaining, even though money is still changing hands.

And while I wish that this phenomenon of "you simply don't have a market" were only relevant to healthcare and not to other facets of our civilization… well, it's not.

vi. Absence of (meta-)competition

VISITOR: I suppose I can imagine a hypothetical world in which *one* country screws things up as badly as you describe. But your planet has multiple governments, I thought. Or did I misunderstand that? Why wouldn't patients emigrate to—or just *visit* —countries that made better hospitals legal?

CECIE: The forces acting on governments with high technology levels are mostly the same between countries, so all the governments of those countries tend to have their medical system screwed up in mostly the same way (not least because they're imitating each other). Some aspects of dysfunctional insurance and payment policies are special to the US, but even the relatively functional National Health System in Britain still has failure of professional specialization. (Though they at least don't require doctors to have philosophy degrees.)

VISITOR: Is there not *one* government that would allow a reasonably designed hospital staffed by specialists instead of generalists?

CECIE: It wouldn't be enough to just have one government's okay. You'd need some way to initially train your workers, despite none of our world's medical schools being set up to train them. A majority of legislators won't benefit *personally* from deciding to let you try your new hospital in their country. Furthermore, you couldn't just go around raising money from rich countries for a venture in a poor country, because rich countries have elaborate regulations on who's allowed to raise money for business ventures through equity sales. The fundamental story is that everything, everywhere, is covered with varying degrees of molasses, and to do any novel thing you have to get around all of the molasses streams *simultaneously*.

VISITOR: So it's impossible to test a functional hospital design *anywhere on the planet*?

CECIE: But of course.

VISITOR: I must still be missing something. I just don't understand why all of the people with economics training on your planet can't go off by themselves and establish their own hospitals. Do you literally have people occupying every square mile of land?

CECIE: … How do I phrase this…

All useful land is already claimed by some national government, in a way that the international order recognizes, whether or not that land is inhabited. No relevant decisionmaker has a personal incentive to allow there to be unclaimed land. Those countries will defend even a very small patch of that claimed land using all of the military force their country has available, and the international order will see you as the aggressor in that case.

VISITOR: Can you *buy* land?

CECIE: You can't buy the sovereignty on the land. Even if you had a *lot* of money, any country poor enough and desperate enough to consider your offer might just steal your stuff after you moved in.

Negotiating the right to bring in weapons to defend yourself in this kind of scenario would be even more unthinkable, and would spark international outrage that could prevent you from trading with other countries.

To be clear, it's not that there's a global dictator who prevents new countries from popping up; but every potentially useful part of every land is under *some* system's control, and all of those systems would refuse you the chance to set up your own alternative system, for very similar reasons.

VISITOR: So there's no way for your planet to *try* different ways of doing things, *anywhere*. You literally cannot run experiments about things like this.

CECIE: Why would there be? Who would decide that, and how would they personally benefit?

VISITOR: That sounds *extremely* alarming. I mean, difficulties of adoption are one thing, but not even being able to *try* new things and see what happens… Shouldn't everyone on your planet be able to detect at a glance how horrible things have become? Can this type of disaster really stand up to *universal* agreement that something is wrong?

CECIE: I'm afraid that our civilization doesn't have a sufficiently stirring and narratively satisfying conception of the valor of "testing things" that our people would be massively alarmed by its impossibility. And now, Visitor, I hope we've bottomed out the general concept of why people can't do things differently—the local system's equilibrium is broken, *and* the larger system's equilibrium makes it impossible to flee the game.

VISITOR: Okay, look… despite everything you've said so far, I still have some trouble understanding why doctors and parents can't just *not* kill the babies. I manage to get up every single morning and successfully not kill any babies. It's not as hard as it sounds.

CECIE: I worry you're starting to think like Simplicio. You can't just *not* kill babies and expect to get away with it.

SIMPLICIO: I actually agree with Cecie here. The evil people behind the system hate those who defy them by behaving differently; there's no way they'd countenance anyone departing from the norm. What we really need is a revolution, so we can depose our corrupt overlords, and finally be free to coordinate, and…!

CECIE: There's no need to add in any evil conspiracy hypotheses here.

It's sufficient to note that the system is *in equilibrium* and it has *causes* for the equilibrium settling there—causes, if not justifications. You can't go against the system's default without going against the forces that underpin that default. A doctor who gives a baby a nutrition formula that isn't FDA-approved will lose their job. A hospital that doesn't fire that kind of doctor will be sued. A scientist that writes proposals for a big, expensive, definitive study won't get a grant, and while they were busy writing those failed grant proposals, they'll have lost their momentum toward tenure. So no, you can't just try out a competing policy of not killing babies. Not more than once.

VISITOR: *Have* you tried?

CECIE: No.

VISITOR: But—

CECIE: Anyway, from my perspective, it's no surprise if you don't yet feel like you understand. We've only *begun* to survey the malfunctions of the whole system, which would further include the FDA, and the clinical trials, and the *p*-hacking. And the way venture capital is structured, and equity-market regulations. And the insurance companies, and the tax code. And the corporations who contract with the insurance companies. And the corporations' employees. And the politicians. And the voters.

VISITOR: … Consider me impressed that your planet managed to reach this level of dysfunction without *actually physically bursting into flames*.

Next: **Moloch's Toolbox** part 2.

The full book will be available November 16th. You can go to equilibriabook.com to pre-order the book, or sign up for notifications about new chapters and other developments.

---

1. Carl Shulman notes that the Affordable Care Act linked federal payments to hospitals with reducing central-line infections (source), which was probably a factor in the change. ↩

2. Around a thousand infants are born with short bowel syndrome per year in the United States, of whom two-thirds develop parenteral nutrition-associated liver disease (source). See Park, Nespor, and Kerner Jr for a 2011 review of the academic literature, and Koch, Cohen, and Carroll and Madrzyk for news coverage. ↩

3. See Tabarrok's "Assessing the FDA via the Anomaly of Off-Label Drug Prescribing," which cites the widespread practice of off-label prescription as evidence that the FDA's efficacy trial requirements are unnecessary. ↩

4. See the "Report Likelihoods, Not *p*-Values" FAQ, or, in dialogue form: "Likelihood Functions, *p*-Values, and the Replication Crisis." ↩

5. From Schmidt and Hunter's "Select on Intelligence": "Intelligence is the major determinant of job performance, and therefore hiring people based on intelligence leads to marked improvements in job performance." See also psychologist Stuart Ritchie's discussion of IQ in *Vox*.

   Software engineer Alyssa Vance adds:

   > I'll note that, as far as I can tell, the informal consensus at least among the best-informed people in software is that hiring has tons of obvious irrationality even when there's definitely no external cause; see [1] and [2]. In terms of Moloch's toolbox, the obvious reason for that is that interviewers are rarely judged on the quality of the people they accept, and when they are, certainly aren't paid more or less based on it. (Never mind the people they reject. "Nobody ever got fired because of the later performance of someone they turned down.") Their incentive, insofar as they have one, is to hire people who they'd most prefer to be on the same floor with all day long. ↩

6. Compare psychiatrist Scott Alexander's account, in "Against Tulip Subsidies":

   > In America, aspiring doctors do four years of undergrad in whatever area they want (I did Philosophy), then four more years of medical school, for a total of eight years post-high school education. In Ireland, aspiring doctors go straight from high school to medical school and finish after five years. I've done medicine in both America and Ireland. The doctors in both countries are about equally good. When Irish doctors take the American standardized tests, they usually do pretty well. Ireland is one of the approximately 100% of First World countries that gets better health outcomes than the United States. There's no evidence whatsoever that American doctors gain anything from those three extra years of undergrad. And why would they? Why is having a philosophy degree under my belt supposed to make me any better at medicine? […]
   >
   > I'll make another confession. Ireland's medical school is five years as opposed to America's four because the Irish spend their first year teaching the basic sciences—biology, organic chemistry, physics, calculus. When I applied to medical school in Ireland, they offered me an accelerated four year program on the grounds that I had surely gotten all of those in my American undergraduate work. I hadn't. I read some books about them over the summer and did just fine.
   >
   > Americans take eight years to become doctors. Irishmen can do it in four, and achieve the same result. Each year of higher education at a good school—let's say an Ivy, doctors don't study at Podunk Community College—costs about $50,000. So American medical students are paying an extra $200,000 for…what?
   >
   > Remember, a modest amount of the current health care crisis is caused by doctors' crippling level of debt. Socially responsible doctors often consider less lucrative careers helping the needy, right up until the bill comes due from their education and they realize they have to make a lot of money right now. We took one look at that problem and said "You know, let's make doctors pay an extra $200,000 for no reason."

   For a more general discussion of the evidence that college is chiefly a costly signal of pre-existing ability, rather than a mechanism for building skills and improving productivity, see Bryan Caplan's argument in "Is College Worth It?", also summarized by Roger Barris. ↩

7. See, e.g., Scott Alexander's "My IRB Nightmare." ↩

8. From Hyman and Silver, "You Get What You Pay For":

   > By the 1950s, death rates ranged between 1 and 10 per 10,000 encounters. Anesthesia mortality stabilized at this rate for more than two decades. Mortality and morbidity rates fell again after a 1978 article reframed the issue of anesthesia safety as one of human factor analysis. In the mid-1980s, the American Society of Anesthesiologists (ASA) promulgated standards of optimal anesthesia practice that relied heavily on systems-based approaches for preventing errors. Because patients frequently sued anesthetists when bad outcomes occurred and because deviations from the ASA guidelines made the imposition of liability much more likely, anesthetists had substantial incentives to comply.

> […] W]e should consider why anesthesia mortality stabilized at a rate more than one hundred times higher than its current level for more than two decades. The problem was not lack of information. To the contrary, anesthesia safety was studied extensively during the period. A better hypothesis is that anesthetists grew accustomed to a mortality rate that was exemplary by health care standards, but that was still higher than it should have been. From a psychological perspective, this low frequency encouraged anesthetists to treat each bad outcome as a tragic but unforeseen and unpreventable event. Indeed, anesthetists likely viewed each individual bad outcome as the manifestation of an irreducible baseline rate of medical mishap.

Hyman and Silver note other possible factors behind the large change, e.g., the fact that the person responsible for mishaps was often easy to identify since there tended to be only one anesthetist per procedure, and that "because surgical patients had no on-going relationships with their anesthetist, victims were particularly likely to sue." ↩

9. See Jena, Prasad, Goldman, and Romley, "Mortality and Treatment Patterns Among Patients Hospitalized With Acute Cardiovascular Conditions During Dates of National Cardiology Meetings." ↩

Modesty and diversity: a concrete suggestion

In online discussions, the number of upvotes or likes a contribution receives is often highly correlated with the social status of the author within that community. This makes the community less epistemically diverse, and can contribute to feelings of groupthink or hero worship.

Yet both the author of a contribution and its degree of support contain bayesian evidenceabout its value, arguably an amount that should overwhelm your own inside view.

We want each individual to invest the socially optimal amount of resources into critically evaluating other people's writing (which is higher than the amount that would be optimal for individual epistemic rationality). Yet we also all and each want to give sufficient weight to authority in forming our all-things-considered views.

As Greg Lewis writes:

> The distinction between 'credence by my lights' versus 'credence all things considered' allows the best of both worlds. One can say 'by my lights, P's credence is X' yet at the same time 'all things considered though, I take P's credence to be Y'. One can form one's own model of P, think the experts are wrong about P, and marshall evidence and arguments for why you are right and they are wrong; yet soberly realise that the chances are you are more likely mistaken; yet also think this effort is nonetheless valuable because even if one is most likely heading down a dead-end, the corporate efforts of people like you promises a good chance of someone finding a better path.

Full blinding to usernames and upvote counts is great for critical thinking. If all you see is the object level, you can't be biased by anything else. The downside is you lose a lot of relevant information. A second downside is that anonymity reduces the selfish incentives to produce good content (we socially reward high-quality, civil discussion, and punish rudeness.)

I have a suggestion for capturing (some of) the best of both worlds:

- first, do all your reading, thinking, upvoting and commenting with full blinding
- once you have finished, un-blind yourself and use the new information to
    - form your all-things-considered view of the topic at hand
    - update your opinion of the people involved in the discussion (for example, if someone was a jerk, you lower your opinion of them).

To enable this, there are now two user scripts which hide usernames and upvote counts on (1) the EA forum and (2) LessWrong 2.0. You'll need to install the Stylish browser extension to use them.

Cross-posted here (clicking the link will unblind you!).

Moloch's Toolbox (2/2)

Follow-up to: Moloch's Toolbox (1/2)

---

vii. Sticky traditions in belief-dependent Nash equilibria without common knowledge

CECIE: I could talk next about a tax system that makes it cheaper for corporations to pay for care instead of patients, and how that sets up a host of "decisionmaker is not the beneficiary" problems.

But I suspect a lot of people reading this conversation understand that part already, so instead I'll turn my attention to venture capital.

VISITOR: It sounds like the "politicians" and the "voters" might be a more key issue, if the cultural translator is right about what those correspond to.

CECIE: Ah! But it turns out that venture capitalists and startups can be seen as a simpler version of voters and politicians, so it's better to consider entrepreneurs first.

Besides, at this point I imagine the Visitor is wondering, "Why can't anyone *make any money* by saving those babies? Doesn't your society have a profit incentive that fixes this?"

VISITOR: Actually, I don't think that *was* high on my list of questions. It's understood among my people that not every problem is one you can make a profit by fixing—*persistent* societal problems tend to be ones that don't have easily capturable profits corresponding to their solution.

I mean, *yes*, if this was all happening on our world and it wasn't already being addressed by the Serious People, then somebody *would* just mix the bleeping nutrients and sell it to the bleeping parents for bleeping money. But at this point I've already guessed that's going to be illegal, or saving babies using money is going to be associated with the wrong Tower and therefore unprestigious, or your parents are using a particular kind of statistical analysis that requires baby sacrifices, or whatever.

CECIE: Hey, details matter!

VISITOR: (*in sad reflection*) Do they? Do they really? Isn't there some point where you just admit you can't stop killing babies and it doesn't really matter why?

CECIE: No. You can *never* say that if you want to go on being a cynical economist.

Now, there are several different kinds of molasses covering the world of startups and venture capital. It's the *tradition-bound* aspects of that ecosystem that we'll find especially interesting, since according to its own ideology, venture capitalists are supposed to chase strange new ideas that other venture capitalists don't believe in. Walking through the simpler case of venture capital will help us understand the more

complex reasons why voters and politicians are nailed into their own equilibria, underpinning the ultimate reasons why nobody can change the laws that prevent change.

VISITOR: (*gazing off into the distance*) … I wonder if maybe there are some worlds that can't be saved.

CECIE: Suppose it's widely believed that the most successful entrepreneurs have red hair. If you're an unusually smart venture capital company that realizes that, *a priori*, hair color doesn't seem like it should correlate to entrepreneurial ability, you might think you could make an excess profit by finding some overlooked entrepreneur with blonde hair.

The key insight here is that venture capital is a *multi-stage* process. There's the initial or pre-seed round, the seed round, the Series A, the Series B, the middle rounds, the Series C… and if the startup fails to raise money on any of those rounds before they become durably profitable, they're dead. What this means is that the seed-round investors need to consider the probability that the company can successfully raise a Series A. If the angels invest in the seed round of a company whose entrepreneurs don't have red hair, that company won't be able to raise a Series A and will go bust and the angel investment will be worthless. So the angel investors need to decide where to invest, and what price to offer, based partially on their beliefs about what most Series A investors believe.

SIMPLICIO: Ah, I've heard of this. It's called a Keynesian beauty contest, where everyone tries to pick the contestant they expect everyone else to pick. A parable illustrating the massive, pointless circularity of the paper game called the stock market, where there's no objective except to buy the pieces of paper you'll think other people will want to buy.

CECIE: No, there are real returns on stocks—usually in the forms of buybacks and acquisitions, nowadays, since dividends are tax-disadvantaged. If the stock market has the nature of a self-fulfilling prophecy, it's only to the extent that high stock prices directly benefit companies, by letting the company get more capital or issue bonds at lower interest. If not for the direct effect that stock prices had on company welfare, it wouldn't matter at all to a 10-year investor what other investors believe today. If stock prices had zero effect on company welfare, you'd be happy to buy the stock that nobody else believed in, and wait for that company to have real revenues and retained assets that everyone else could see 10 years later.

SIMPLICIO: But *nobody* invests on a 10-year horizon! Even pension companies invest to manage the pension manager's bonus this year!

VISITOR: Surely the recursive argument is obvious? If most managers invest with 1-year lookahead, a smarter manager can make a profit in 1 year by investing with a 2-year lookahead, and can continue to extract value until there's no predictable change from 2-year prices to 1-year prices.

CECIE: In the entrepreneurial world, startups are killed outright, very quickly, by the equivalent of low stock prices. *And* for legal reasons there are no hedge funds that can adjust market prices en masse, so the recursive argument doesn't apply. The upshot is that seed investors have a *strong* incentive to care about what Series A investors think. If the entrepreneurs don't fit the stereotype of cool entrepreneurs who have red hair, you can't make an excess return by going against the popular misapprehension, because the startup will die in the next funding round.

The key phenomenon underlying the social molasses is that there's a self-reinforcing equilibrium of beliefs. Maybe a *lot* of the Series A investors think the idea of entrepreneurs needing to have red hair is objectively silly. But they expect Series B investors to believe it. So the Series A investors don't invest in blonde-haired entrepreneurs. So the seed investors are right to believe that "Series A investors won't invest in blonde-haired companies" even if a lot of the reason why Series A investors aren't investing is not that they believe the stereotype but that they believe that Series B investors believe the stereotype. And from the outside, of course, all that investors can *see* is that most investors aren't investing in blonde-haired entrepreneurs—which just goes to reinforce everyone's belief that everyone else believes that red-haired entrepreneurs do better.[10]

VISITOR: And you can't just have everyone say those exact words aloud, in unison, and simultaneously wake up from the dream?

SIMPLICIO: I'm afraid people don't understand recursion as well as that would require.

CECIE: Perhaps, Simplicio, it is only that most VCs believe that most other VCs don't understand recursion; that would have much the same effect in practice.

SIMPLICIO: Or maybe most people *are* too stupid to understand recursion. Is that something you'd be able to accept, if it were true?

CECIE: Regardless, on a larger scale, what we're seeing is an extra stickiness that results when the incentive to try an innovation requires you to believe that *other people will believe* the innovation will work. An equilibrium like that can be *much stickier* than a scenario where, if *you* believe that a project will succeed, *you* have an incentive to try it *even if* other people expect the project to fail.

Stereotypically, the startup world is supposed to consist of heroes producing an excess return by pursuing ideas that nobody else believes in. In reality, the multi-stage nature of venture capital makes it very easy for the field to end up pinned to traditions about whether entrepreneurs ought to have red hair—not because everyone believes it, but because everyone believes that everyone believes it.

viii. First-past-the-post and wasted votes

VISITOR: Does this feed back into our primary question of why your society can't stop itself from feeding poisonous substances to babies?

CECIE: It's true that venture capitalists are now collectively skeptical of attempts at new drug development, but the real problem (at least for cases like this) is the enormous cost of approval and the long delays the FDA causes.[11] The actual reason I went into this is that by understanding venture capitalists and entrepreneurs, we can understand the more complex case of voters and politicians. Which is the key to the *political* equilibrium that pins down the FDA, and all the other laws that prevent anyone from doing better. Not always, but quite often, the ultimate foundations of failure trace back to the molasses covering voters and politicians.

SIMPLICIO: I'd like to offer, throughout whatever theory follows, the alternative hypothesis that voters are *in fact* just fools, sheep, and knaves. I mean, you should at

least be considering that possibility.

CECIE: The simplest way of understanding the analogy between venture capitalists and voters is that voters have to vote for politicians that are electable.

VISITOR: Uh, what? When you write down your preference ordering on elected representatives, you need to put politicians that other voters prefer at the top of your preference ordering?

CECIE: Yes, that's pretty much what it amounts to. In the US, at least, elections are run on what's known as a "first-past-the-post" voting system. Whoever gets the most votes in the contest wins. People who study voting systems widely agree that first-past-the-post is among the *worst* voting systems—it's provably impossible for one voting system to have all the intuitively good properties at once, but FPTP is one of the *most* broken.

VISITOR: Why not vote to change the voting system, then?

CECIE: I'll get to that!

There are several ways of explaining what's wrong with FPTP, but a lovely explanation I recently encountered phrases the explanation in terms of "wasted votes"—the total number of votes that can be removed without changing the outcome.

The two classic forms of gerrymandering are *cracking* and *packing*. Let's say the parties are Green and Orange, and the Green party is in charge of drawing the voting boundaries. As a Green, you want to draw up districts such that Green politicians win with 55% of the vote—with some room for error, but not all that much—and for Orange politicians to win with 100% of the vote.

SIMPLICIO: Ah, so that the Orange politicians won't need to be responsive to Orange voters because their re-election is nearly guaranteed, right?

CECIE: No, the plot is far more diabolical than that. Consider a district of 100,000 people, where a Green politician wins with 55% of the vote. When 50,001 Green voters had cast their ballots, the election was already decided, under first-past-the-post, so the next 4,999 Green votes are "wasted"—this is to be understood as a technical term, not a moral judgment—in that they don't further change the outcome. Then 45,000 Orange votes are also "wasted," in that they don't change the outcome. And also, one notes, those Orange voters don't get the representative they wanted.

In an Orange district of 100,000 where the politician wins with 100% of the vote, there are 50,000 potent Orange votes and 50,000 wasted Orange votes. In total, there are 50,000 potent Green votes, 5,000 wasted Green votes, 50,000 potent Orange votes, and 95,000 wasted Orange votes. On a larger scale, this means that you can control a majority of a state legislature with slightly more than 1/4 of the votes—just have 55% of the districts containing 55% Green voters, with everything else solid Orange.

VISITOR: And then this quarter of the population rules cruelly over the remaining three-quarters, who in turn lack the weapons to rise up?

CECIE: No, the real damage is far subtler. Let's say that Alice, Bob, and Carol have taken time off from their cryptographic shenanigans to run for political office. Alice is in the lead, followed by Bob and then by Carol. Suppose Dennis prefers Carol to Bob, and Bob to Alice. But Dennis can't actually write "Carol > Bob > Alice" on a slip of

paper that gets processed by a trivially more sophisticated voting system. Dennis is only allowed to write down one candidate's name, and that's his vote. Under a system where the candidate with the most votes wins, and there's uncertainty about which of the two frontrunners might win, *all votes for whoever is in third place will be wasted votes*, and this fact is predictable to the voters.

VISITOR: Ah, I see. That's why you introduced your peculiar multi-stage system of venture capital, which I assume must be held in place by laws forbidding anyone else to go off and organize their own financial system differently, and observed how it creates a sticky equilibrium in which financiers must believe that other financiers will believe in a startup.

If Dennis doesn't believe that other "voters" will believe in Carol, Dennis will vote for Bob, which makes your politics stickier than a system in which "voters" were permitted to support the people they actually liked.

CECIE: Well, you see the analogy, but I'm not sure you appreciate the true depth of the horror.

VISITOR: I'm sure I don't.

CECIE: The upshot of first-past-the-post is typically a political system dominated by exactly two parties.

VISITOR: Parties?

SIMPLICIO: Entities that tell sheep who to vote for.

CECIE: In elections that have a single winner, votes for any candidate who isn't one of the top two choices are wasted. In a representative democracy where districts vote on representatives who vote on laws, the dynamics of the district vote are then influenced by the dynamics of the national vote. Even if a third-party candidate could win a district, they wouldn't have anyone to work with in the legislature, and so their votes would generally be wasted.

In the absence of a way to solve a large coordination problem, there's no way for a third party to gain marginal influence over time. Each individual who considers voting for a third-party candidate knows they'll be wasting their vote. This also means that third parties can't field good candidates, since potential candidates know they'd be running to lose, which is stressful and unrewarding for people with better life options. And that's a sufficient multi-factor system to prevent strong third parties from arising. When you're not allowed to vote for Carol, who you actually like, you'll vote for whichever of Alice and Bob you dislike the least.

The resulting equilibrium… well, Abramowitz and Webster found that what mainly predicted voting behavior wasn't how much the voter liked their preferred party, but how much they disliked the opposing party.[12] Essentially, the US has two major voting factions, "people who hate Red politicians" and "people who hate Blue politicians." When the Red politicians do something that Red-haters *really* dislike, that gives the Blue politicians more leeway to do additional things that Red-haters mildly dislike, which can give the Red politicians more leeway of their own, and so the whole thing slides sideways.

SIMPLICIO: Looking at the abstract of that Abramowitz and Webster paper, isn't one of their major findings that this type of hate-based polarization has *increased* a great

deal over the last twenty years?

CECIE: Well, yes. I don't claim to know exactly why that happened, but I suspect the Internet had something to do with it.

In the US, the current two parties froze into place in the early twentieth century—before then, there was sometimes turnover (or threatened turnover). I suspect that the spread of radio broadcasting had something to do with the freeze. If you imagine a country in the pre-telegraph days, then it might be possible for third-party candidates to take hold in one state, then in nearby states, and so a global change starts from a local nucleus. A national radio system makes politics less local.

The Internet might have pushed this phenomenon further and caused most of politics to be about the same national issues, which in turn reinforces the Red-vs.-Blue dynamic that allows each party to sustain itself on hatred for the other.

But that's just me trying to eyeball the phenomenon using American history—I haven't studied it. Other countries that also have the radio and Internet and similar electoral dynamics do manage to have more than two relevant parties, possibly because of dynamics that cause the votes of third-party politicians to be less wasted.

SIMPLICIO: Isn't the solution here obvious, though? All of these problems are caused by voters' willingness to compromise on their principles and accept the lesser of two evils.

CECIE: Would things be better if people chose the greater of two evils? If they acted ineffectually against that greater evil? The Nash equilibrium isn't an illusion. Individuals would do worse by playing away from that Nash equilibrium. Wasted votes *are* wasted. The current system *is* an effective trap and the voters *are* trapped. They can't just wish their way out of that trap.

There doesn't need to be any way for good to win; and if there isn't, the lesser evil really is the best that voters can do. Pretending otherwise may feel righteous, but it doesn't change the equilibrium.

VISITOR: Just one second. Isn't this all window dressing, compared to the issue of whatever true ruler imposes these rules on the "voters"? Like, if you put me into an elaborate cage that gives me an electric shock each time I vote for Carol, obviously the person who really controls the system is whoever put the cage in place and determines which politicians you can vote for without electric shocks.

SIMPLICIO: I like the way you think.

CECIE: It's not *quite* true to say that the system is self-reinforcing and that the voters are the sole instrument of their own destruction. But the lack of any obvious, individual tyrant who personally decides who you're allowed to vote for has indeed caused many voters to believe that they are in control. I mean, they don't *feel* like they're in control, but they think that "the voters" select politicians.

They aren't able to personalize a complicated bad equilibrium as a tyrant—not like they would blame a jeweled king who was standing in the polling booth, ready to give them an electric shock if they wrote down Carol's name.

Inspired by Allan Ginsberg's poem *Moloch*, Scott Alexander once wrote of coordination failures:

> Moloch is introduced as the answer to a question—C. S. Lewis' question in Hierarchy Of Philosophers—*what does it*? Earth could be fair, and all men glad and wise. Instead we have prisons, smokestacks, asylums. What sphinx of cement and aluminum breaks open their skulls and eats up their imagination?
>
> And Ginsberg answers: *Moloch does it.*
>
> There's a passage in the *Principia Discordia* where Malaclypse complains to the Goddess about the evils of human society. "Everyone is hurting each other, the planet is rampant with injustices, whole societies plunder groups of their own people, mothers imprison sons, children perish while brothers war."
>
> The Goddess answers: "What is the matter with that, if it's what you want to do?"
>
> Malaclypse: "But nobody wants it! Everybody hates it!"
>
> Goddess: "Oh. Well, then stop."
>
> The implicit question is—if everyone hates the current system, who perpetuates it? And Ginsberg answers: "Moloch." It's powerful not because it's correct—nobody literally thinks an ancient Carthaginian demon causes everything—but because thinking of the system as an agent throws into relief the degree to which the system *isn't* an agent.[13]

Scott Alexander saw the face of the Enemy, and he gave it a name—thinking that perhaps that would help.

VISITOR: So if you did do this to yourselves, all by yourselves with no external empire to prevent you from doing anything differently by force of arms, then *why can't you just vote to change the voting rules*? No, never mind "voting"—why can't you all just *get together and change everything, period*?

CECIE: It's true that concepts like these are nontrivial to understand.

It's not obvious to me that people *couldn't possibly* understand them, if somebody worked for a while on creating diagrams and videos.

But the bigger problem is that people wouldn't know they could trust the diagrams and videos. I suspect some of the dynamics in entrepreneur-land are there because many venture capitalists run into entrepreneurs that are smarter than them, but who still have bad startups. A venture capitalist who believes clever-sounding arguments will soon be talked into wasting a lot of money. So venture capitalists learn to distrust clever-sounding arguments because they can't distinguish lies from truth, when they're up against entrepreneurs who are smarter than them.

Similarly, the average politician is smarter than the average voter, so by now most voters are just accustomed to a haze of plausible-sounding arguments. It's not that you can't possibly explain a Nash equilibrium. It's that there are too many people advocating changes in the system for their own reasons, who could also draw diagrams that sounded equally convincing to someone who didn't already understand Nash equilibria. Any talk of systemic change on this level would just be lost in a haze of equally plausible-sounding-to-the-average-voter blogs, talking about how quantitative easing will cause hyperinflation.

VISITOR: Maybe it's naive of me… but I can't help but think… that *surely* there must be *some* breaking point in this system you describe, of voting for the less bad of two awful people, where the candidates just get worse and worse over time. At some point, shouldn't this be trumped by the "voters" just getting completely fed up? A spontaneous equilibrium-breaking, where they just didn't vote for either of the standard lizards no matter what?

CECIE: Perhaps so! But my own cynicism can't help but suspect that this "trumping" phenomenon of which you speak would be even worse.

SIMPLICIO: I have a technical objection to your ascribing all these sins to first-past-the-post voting rather than, say, the personal vices of the voters. There are numerous parliamentary democracies outside the United States that practice proportional representation, where a party getting 30% of the votes gets 30% of the seats in parliament. And *they* don't seem to have solved these problems.

CECIE: Omegaven does happen to be approved in Europe, however. Like, they are not in fact killing those particular babies—

SIMPLICIO: Oh, *come on!* Yes, the European equivalent of the US's FDA happens to be a bit less stupid. Lots of other things in European countries happen to be more stupid. Indeed, I'd say that in Europe you have much *crazier* people getting seats in parliaments, compared to the United States. The problem isn't the voting system. The problem is the *voters*.

CECIE: There are indeed some voters who want stupid things, and under the European system, their voice can be heard. There are also voters who want smart things and whose voices can be heard, like in the Pirate Party in Finland. But European parliamentary systems have *different* problems stemming from *different* systemic flaws.

Proportional representation would be a good system for a legislature that needed to repeatedly vote on laws, where different legislators could form different coalitions for each vote. If instead you demand that a majority coalition "form a government" to appoint an executive, then you need to give concessions to some factions, while other factions get frozen out. I'm not necessarily saying that it would be *easy* to fix all the problems simultaneously. Still, I imagine that a proportionally represented legislature, *combined* with an executive elected at-large by Condorcet voting, might possibly be less stupid—

SIMPLICIO: Or maybe it would just give stupid voters a louder voice. I don't like the evil conspiracy of the press and political elites that governs my country from the shadows, but I *am* willing to consider the proposition that the alternative is Donald Trump. I mean, I intend to go on fighting the Conspiracy about many specific issues. But if you're proposing a reform that puts more power into the hands of sheep not yet awakened, the results could be even worse.

CECIE: Well, I agree that the design of well-functioning political systems is hard. Singapore might be the best-governed country in the world, and their history is approximately, "Lee Kuan Yew gained very strong individual power over a small country, and unlike the hundreds of times in the history of Earth when that went horribly wrong, Lee Kuan Yew happened to know some economics." But the Visitor asked me why we were killing babies, and I tried to answer in terms of the system that obtained in the part of the world that was actually killing those babies. *You* asked why Europe wasn't a paradise since it used proportional representation, and my answer is

that parliamentary systems have their own design flaws that induce a different kind of dysfunction.

SIMPLICIO: Then if both systems are bad, how does your hypothesis have any observable consequences?

CECIE: Because different systems are bad in different ways. When you have a "crazy" new idea, whether it's good or bad, the European parliaments will be allowed to talk about it first. Whether that's Omegaven, basic income, gay marriage, legalized prostitution, ending the war on drugs, land value taxes, or fascist nationalism, you are more likely to find it talked about in systems of proportional representation. It also happens to be true that those governments bloat up faster because of the repeated bribes required to hold the "governing coalition" together, but that's a *different* problem.

ix. The Overton window

SIMPLICIO: I'm beginning to experience the same sort of confusion as the Visitor about your view of the world, Conventional Cynical Economist. If voters weren't stupid, the world would look very different than it does.

If the ultimate source of stupidity were poorly designed governmental structures, then average voters would sound smarter than average politicians. I don't think that's *actually* true.

CECIE: There are deeper forms of psychological molasses that generalize beyond first-past-the-post political candidates. The still greater force locking bad political systems into place is an equilibrium of silence about policies that aren't "serious."

A journalist thinks that a candidate who talks about ending the War on Drugs isn't a "serious candidate." And the newspaper won't cover that candidate because the newspaper itself wants to look serious… or they think voters won't be interested because everyone knows that candidate can't win, or something? Maybe in a US-style system, only contrarians and other people who lack the social skill of getting along with the System are voting for Carol, so Carol is uncool the same way Velcro is uncool and so are all her policies and ideas? I'm not sure exactly what the journalists are thinking subjectively, since I'm not a journalist. But if an existing politician talks about a policy outside of what journalists think is appealing to voters, the journalists think the politician has committed a gaffe, and they write about this sports blunder by the politician, and the actual voters take their cues from that. So no politician talks about things that a journalist believes it would be a blunder for a politician to talk about. The space of what it isn't a "blunder" for a politician to talk about is conventionally termed the "Overton window."

SIMPLICIO: It's all well and good to talk about complicated clever things, Cynical Economist, but what explanatory power does all this added complexity have? Why postulate politicians who believe that journalists believe that voters won't take something seriously? Why not just say that people are sheep?

CECIE: To name a recent example from the United States, it explains how, one year, gay marriage is this taboo topic, and then all of a sudden there's a huge upswing in

everyone being *allowed* to talk about it for the first time and shortly afterwards it's a done deal. If you suppose that a huge number of people really did hate gay marriage deep down, or that all the politicians mouthing off about the sanctity of marriage were engaged in a dark conspiracy, then why the sudden change?

With my more complicated model, we can say, "An increasing number of people over time thought that gay marriage was pretty much okay. But while that group didn't have a majority, journalists modeled a gay marriage endorsement as a 'gaffe' or 'unelectable', something they'd write about in the sports-coverage overtone of a blunder by the other team—"

SIMPLICIO: Ah, so you say it was a conspiracy by evil journalists?

CECIE: No! Those journalists weren't *consciously deciding* the equilibrium. The journalists were writing "serious" articles, i.e., articles about Alice and Bob rather than Carol. The equilibrium *consisted of* the journalists writing sports coverage of elections, where everything is viewed through the lens of a zero-sum competition for votes between Alice's team and Bob's team. Viewed through that lens, the journalists thought a gay marriage endorsement would be a blunder. And if you do something that enough journalists think is a political blunder, it *is* a political blunder. The journalists' sports coverage will describe you as an incompetent politician, and primates instinctively want to ally with likely winners. Which meant the equilibrium could have a sharp tipover point, *without* most of the actual population changing their minds sharply about gay marriage in that particular year. The support level went over a threshold where somebody tested the waters and got away with it, and journalists began to suspect it wasn't a political blunder to support gay marriage, which let more politicians speak and get away with it, and then the *change of belief about what was inside the Overton window* snowballed. I think that's what we saw.

SIMPLICIO: Forgive me for resorting to Occam's Razor, but is it not simpler just to say that people's beliefs changed slowly until it reached some level where the military-industrial complex realized they couldn't win the battle to suppress gay marriage outright, and so stopped fighting?

CECIE: In a sense, that's not far off from what happened, except without the evil conspiracy part. We might or might not be approaching a similar tipover point about ending the War on Drugs—a long, slow, secular shift in opinion, followed by a sudden tipover point where journalists model politicians as being allowed to talk about it, which means that politicians *can* talk about it, and then a few years later everyone is acting like they always thought that way. At least, I *hope* that's where the current trend is leading.

SIMPLICIO: Several states have already passed laws legalizing marijuana. Why hasn't that already broken the Overton window?

CECIE: Because voter initiatives don't break the common belief about what it would be a "gaffe" for a *serious, national-level* politician to do.

ELIEZER: (*aside*) What broke the silence about artificial general intelligence (AGI) in 2014 wasn't Stephen Hawking writing a careful, well-considered essay about how this was a real issue. The silence only broke when Elon Musk tweeted about Nick Bostrom's *Superintelligence*, and then made an off-the-cuff remark about how AGI was "summoning the demon."

Why did that heave a rock through the Overton window, when Stephen Hawking couldn't? Because Stephen Hawking *sounded like* he was trying hard to appear sober and serious, which signals that this is a subject you have to be careful not to gaffe about. And then Elon Musk was like, "*Whoa, look at that apocalypse over there!!*" After which there was the equivalent of journalists trying to pile on, shouting, "A gaffe! A gaffe! A… gaffe?" and finding out that, in light of recent news stories about AI and in light of Elon Musk's good reputation, people weren't backing them up on that gaffe thing.

Similarly, to heave a rock through the Overton window on the War on Drugs, what you need is not state propositions (although those do help) or articles in *The Economist*. What you need is for some "serious" politician to say, "This is dumb," and for the journalists to pile on shouting, "A gaffe! A gaffe… a gaffe?" But it's a grave personal risk for a politician to test whether the public atmosphere has changed enough, and even if it worked, they'd capture very little of the human benefit for themselves.

VISITOR: So… if this is the key meta-level problem… then why can't your civilization just *consider and solve this entire problem on the meta level*?

CECIE: Oh, I'm afraid that this entire meta-problem isn't the sort of thing the "leading candidates" Alice and Bob talk about, so the problem itself isn't viewed as serious. That is, journalists won't think it's serious. Meta-problems in general—even problems as simple as first-past-the-post versus instant runoff for particular electoral districts—are issues outside the Overton window. So the leading candidates Alice and Bob won't talk about organizational design reform, because it would be very damaging to their careers if they visibly focused their attention on issues that journalists don't think of as "serious."

VISITOR: Then perhaps the deeper question is, "Why does anyone listen to these 'journalists'?" You keep attributing power to them, but you haven't yet explained why they have that power under your equilibrium.

CECIE: People believe that other people believe what's in the newspapers.

Well, no, that's too optimistic. A lot of people *do* believe what's in the newspapers, so long as it isn't about a topic regarding which they have any personal knowledge or expertise. The Gell-Mann Amnesia Effect is the term for how we read the paper about subjects we know about, and it's talking about how wet streets cause rain; and then we turn to the story about international affairs or dieting, and for some reason assume it's more accurate.

There's some level on which most people prefer to talk and believe within the same mental world as other people. Nowadays a lot of people believe what they read on, say, Tumblr, and hardly look at *The New York Times* at all. But even then they still believe that *other people* believe what's in *The New York Times*. That's what gives *The New York Times* its special power over the collective consciousness, far out of proportion to their dwindling readership or the vanishing real trust that individuals from various walks of life have in them—what's printed in *The New York Times* determines what people believe other people believe.

SIMPLICIO: Do you *truly* lay all the sins of humanity at the feet of all this weird recursion? Or is this just a sufficiently weird hypothesis that you find it more fun to think about than the alternatives?

CECIE: I'm not sure I'm pointing in exactly the right direction, but I feel that I'm pointing in the general direction of something that's truly important to the Visitor's most basic question. The Visitor keeps asking why, in some sense, on some sufficiently general level, we can't *just snap out of it*. And to put it in the sort of terms you yourself might want to use, Simplicio, if we're looking for an explanation of why we can't *just snap out of it*, then it might make sense to point to a bad Nash equilibrium covering our collective consciousness and discussion. I suspect that the recursion, the dependency on what people believe other people believe, has a lot to do with making that a *sticky* equilibrium a la venture capital.

ELIEZER: (*aside*) Returning to my day job: As of 2017, I pretty commonly hear from AI researchers who are worried about AGI safety, but who say that they don't dare say anything like that aloud. You could see this as either a good sign or a very bad sign, depending on how pessimistic or optimistic you previously were about the adequacy of academic discussion.

SIMPLICIO: But then what, on your view, is the better way?

CECIE: Again, I could pontificate about various ideas, but that's a *different and harder question* than looking at the actual equilibrium that currently obtains and forces doctors to poison babies. There doesn't have to be a better way.

x. Lower-hanging altruistic fruit and bigger problems

(The Visitor takes a deep breath. When the Visitor speaks again, it is louder.)

VISITOR: Then what about your <*untranslatable 17*>?

CECIE: Sorry? That word didn't come through.

VISITOR: What about everyone on your *entire planet* who could possibly care about babies dying?

So your medical specialists are borked. From the magic-tower analogy, I assume your systems of learning are borked, and that means most of the parents whose responsibility it is to protect the child are borked. Your politicians are borked. Your voters are borked. Your planet has no Serious People who could be trusted to try alternative shoe designs, let alone lead the way on any more complex coordination problem. Your prediction markets, I suppose, are somehow borked in a way that prevents anyone from making a profit by correcting inaccurate policy forecasts… maybe they forecast wrongly bad consequences to unpopular policies, which therefore never get implemented in a way that shows up the inaccurate prediction, since you don't have any way to test things on a smaller scale? Your economists must somehow be borked—

CECIE: It's more that nobody ever listens to us. They *pay us* and then they don't *listen to us*.

VISITOR: —and your financial system is borked so that nobody can make a profit on saving those babies or doing anything else useful. I'm not stupid. I've picked up on the pattern at this point.

But what about *everyone else?* There are seven *billion* people on your planet. How is it that *none* of them step up to save these babies from death and brain damage? How is your *entire planet* failing to solve this problem?

CECIE: That… sounds like a weird question, to an Earth person.

VISITOR: Whatever your problems are, surely out of seven billion human beings there have to be *some* who could *see* the problems as you've laid them out, who could try to rally others to the cause of saving those babies, who could do *whatever it took* to save them!

Even if your system declares that saving babies is only the responsibility of "doctors" or "politicians" or whoever is the Someone Else whose Problem it is, there's no law of physics that *stops* someone else from walking up to the problem and accepting responsibility for it. Out of seven billion people in your world, I can't believe that *literally all* of them are incapable of gathering together some friends and starting things down the path to getting a little fish oil into a baby's nutritional mixture!

ELIEZER: I think I'll step in myself at this point. There's one other very general conclusion we can draw from seeing this ever-growing heap of dead babies. We might say, "the inadequacy of the part implies the inadequacy of the whole"—as we've defined our terms, if a part of the system is inadequate in X lives saved for Y dollars, then the whole system is inadequate in X lives saved for Y dollars. Someone who is motivated and maximizing will first go after the biggest inadequacy *anywhere* that they think they can solve, and if they succeed, it pushes forward the adequacy frontier for the whole system. Thus, we can draw one other general conclusion from the observation that babies are still being fed soybean oil. We can conclude that everyone on the planet who is smart enough to understand this problem, and who cares about strangers' lives, and who maximizes over their opportunities, must have *something more important to do* than getting started on solving it.

VISITOR: (*aghast*) More important than saving hundreds of babies per year from dying or suffering permanent brain damage?

ELIEZER: The observation stands: there must be, in fact, literally nobody on Earth who can read Wikipedia entries and understand that omega-6 and omega-3 fats are different micronutrients, who also cares and maximizes and can head up new projects, who thinks that saving a few hundred babies per year from death and permanent brain damage is the most important thing they could do with their lives.

VISITOR: So you're implying…

ELIEZER: Well, mostly I'm implying that *maximizing* altruism is incredibly rare, especially when you also require sufficiently precise reasoning that you aren't limited to cases where the large-scale, convincing study has already been done; and then we're demanding the executive ability to start a new project on top of that. But yes, I'm also saying that here on Earth we have much more horrible problems to worry about.

CECIE: We've just been walking through a handful of lay economic concepts here, the kind whose structure I can explain in a few thousand words. If you truly perceived the world through the eyes of a conventional cynical economist, then the horrors, the abominations, the low-hanging fruits you saw unpicked would annihilate your very soul.

VISITOR: …

ELIEZER: And then some of us have much, *much* more horrible problems to worry about. Problems that take *more* than reading Wikipedia entries to understand, so that the pool of potential solvers is even smaller. But even just considering this particular heap of dead babies, we know from observation that this part must be true: If you imagine everyone on Earth who fits the qualifications for the dead-baby problem—enough scientific literacy to understand relevant facts about metabolic pathways, *and* the caring, *and* the maximization, *and* enough scrappiness to be the first one who gets started on it, meeting in a conference room to divide up Earth's most important problems, with the first subgroup taking on the most neglected problems demanding the most specialized background knowledge, and the second taking on the second-most-incomprehensible set of problems, until the crowdedness of the previously most urgent problem decreases the marginal impact of further contributions to the point where the next-worst problem at that level of background knowledge and insight becomes attractive… and so on down the ladders of urgency inside the levels of discernment… then there must be such a long and terrible list of tasks left undone, and so few people to understand and care, that saving a few hundred babies per year from dying or suffering permanent brain damage didn't make the list. So it has been observed, and so it must be.

WANDERING BYSTANDER: (*interjecting*) But I just can't believe our planet would be that dysfunctional. Therefore, by backward chaining, I question the original observation on which you founded your inference. In particular, I'm starting to wonder whether omega-3 and omega-6 could *really* be such significantly different micronutrients. Maybe that's just a crackpot diet theory that somehow made it into Wikipedia, and actually all fats *are* pretty much the same, so there's nothing especially terrifying about the prospect of feeding babies exclusively fat from soybean oil instead of something more closely resembling the lipid profile of breast milk?

ELIEZER: Ah, yes. I'm glad you spoke up. I'll get to your modest proposal next.

Next: **Living in an Inadequate World.**

---

10. See Glenn Loury's *The Anatomy of Racial Inequality* for an early discussion of this issue. Note that some venture capitalists I've spoken to endorse this as an account of VC dysfunction, while others have different hypotheses. ↩

11. Carl Shulman argues that the FDA's clinical trial requirements probably *aren't* the reason for recent decades' slowdown in the development of cool new drugs, given that increased regulation seems to have coincided with but not substantially accelerated the declining efficiency of pharmaceutical research and development (source). Shulman suggests that Baumol's cost disease and diminishing returns play a larger role in the R&D slowdown.

    The FDA's clinical trial requirements are much more likely to play a central role in limiting access to non-patented substances, though it's worth noting here that the FDA has gotten faster than it used to be (source). ↩

12. Abramowitz and Webster, "All Politics is National." ↩

13. See Scott Alexander's "Meditations on Moloch." ↩

The Copernican Revolution from the Inside

The Copernican revolution was a pivotal event in the history of science. Yet I believe that the lessons most often taught from from this period are largely historically inaccurate and that the most important lessons are basically *not taught at all* [1]. As it turns out, the history of the Copernican revolution carries important and surprising lessons about rationality -- about what it is and is not like to figure out how the world actually works. Also, it's relevant to deep learning, but it'll take me about 5000 words on renaissance astronomy to make that point.

I used to view the Copernican revolution as an epic triumph of reason over superstition, of open science over closed dogma. Basically, things went as follows: Copernicus figured out that the sun rather than the earth is at the center of our planetary system. This theory immediately made sense of the available data, undermining its contorted predecessors with dazzling elegance. Yet its adoption was delayed by the Catholic Church fighting tooth and claw to keep the truth at bay. Eventually, with the emergence of Newton's work and the dawn of the Enlightenment, heliocentrism became undeniable and its adoption inevitable [2].

This view is inaccurate. Copernicus system was *not* immediately superior. It was rejected by many people who were *not* puppets of the Church. And among those who did accept it, better fit to the data was *not* a main reason. What did in fact happen will become clear in a moment. But in reading that, I'd like to prompt you to consider the events from a very particular vantage point: namely what they would be like *from the inside*. Ask yourself not what these events seem like for a millennial with the overpowered benefit of historical hindsight, but for a Prussian astronomer, an English nobleman or a Dominican priest.

More precisely, there are two key questions here.

First, if you lived in the time of the Copernican revolution, would you have accepted heliocentrism? I don't mean this as a social question, regarding whether you would have had the courage and resources to stand up to the immensely powerful Catholic Church. Rather, this is an epistemic question: based on the evidence and arguments available to you, would you have accepted heliocentrism? For most of us, I think the answer is unfortunately, emphatically, and surprisingly, *no*. The more I've read about the Copernican revolution, the less I've viewed it as a key insight followed by a social struggle. Instead I now view it as a complete mess: of inconsistent data, idiosyncratic mysticism, correct arguments, equally convincing arguments *that were wrong*, and various social and religious struggles thrown in *as well*. It seems to me an incredibly valuable exercise to try and feel this mess from the inside, in order to gain a sense what intellectual progress, historically, has actually been like. Hence a key reason for writing this post is not to provide any clear answers -- although I will make some tentative suggestions -- but to provoke a legitimate sense of confusion.

If things were that chaotic, then this raises the second question. How should you develop intellectually, in order to become the kind of person who would have accepted heliocentrism during the Copernican revolution? Which intellectual habits, if any, unite heliocentric thinkers like Copernicus, Kepler, Galileo and Descartes, and separates them from thinkers like Ptolemy and Tycho? Once again, my answer will be tentative and limited. But my questions, on the other hand, are arguably the right ones.

What happened

My view of the Copernican revolution used to be that when people finally switched to the heliocentric model, *something clicked*. The data was suddenly predictable and

understandable. Something like how Andrew Wiles describes his experience of doing mathematics:

“[...] in terms of entering a dark mansion. You go into the first room and it's dark, completely dark. You stumble around, bumping into the furniture. Gradually, you learn where each piece of furniture is. And finally, after six months or so, you find the light switch and turn it on. Suddenly, it's all illuminated and you can see exactly where you were.”

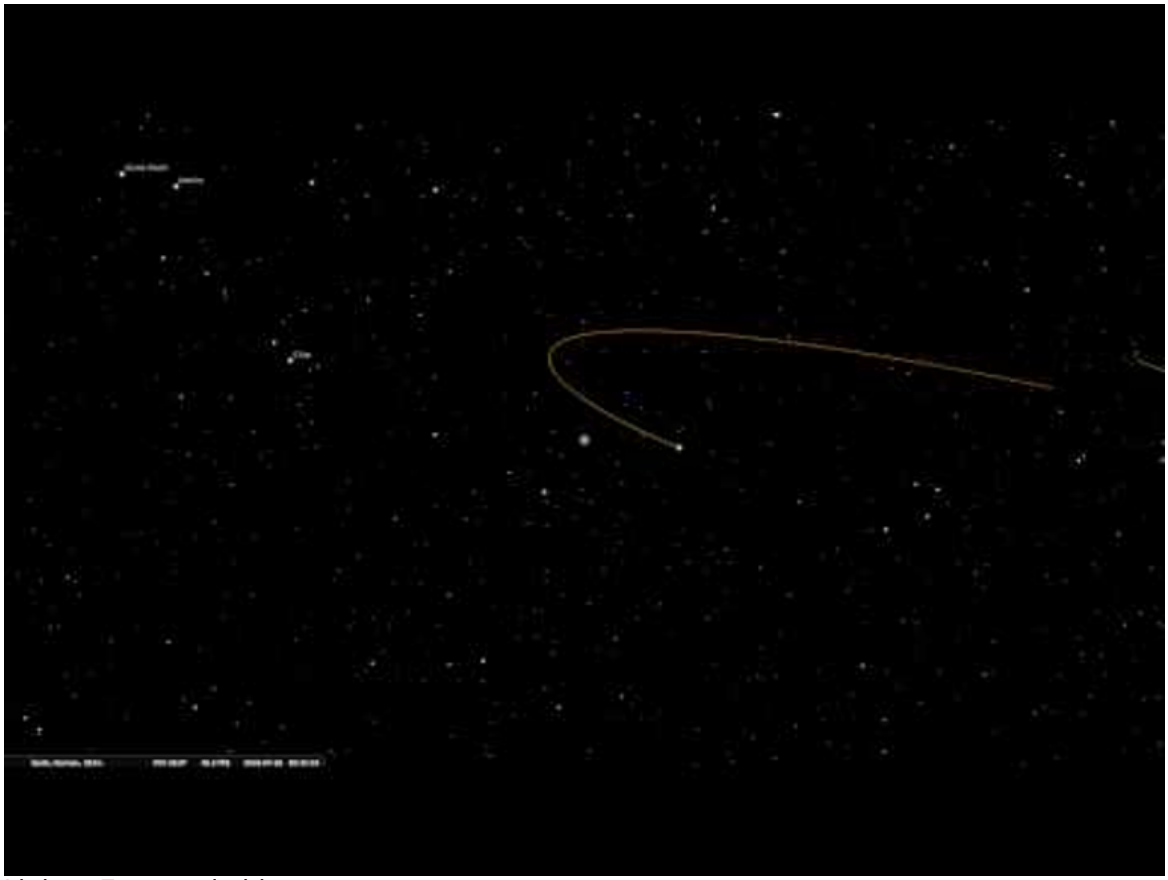
However, this is most certainly *not* how things appeared at the time. Let’s start at the beginning.

1. Scholasticism

The dominant medieval theory of physics, and by extension astronomy, was Scholasticism, a combination of Aristotelian physics and Christian theology. Scholasticism was a geocentric view. It placed the earth firmly at the center of the universe, and surrounded it with a series of concentric, rotating “crystalline spheres”, to which the celestial bodies were attached.

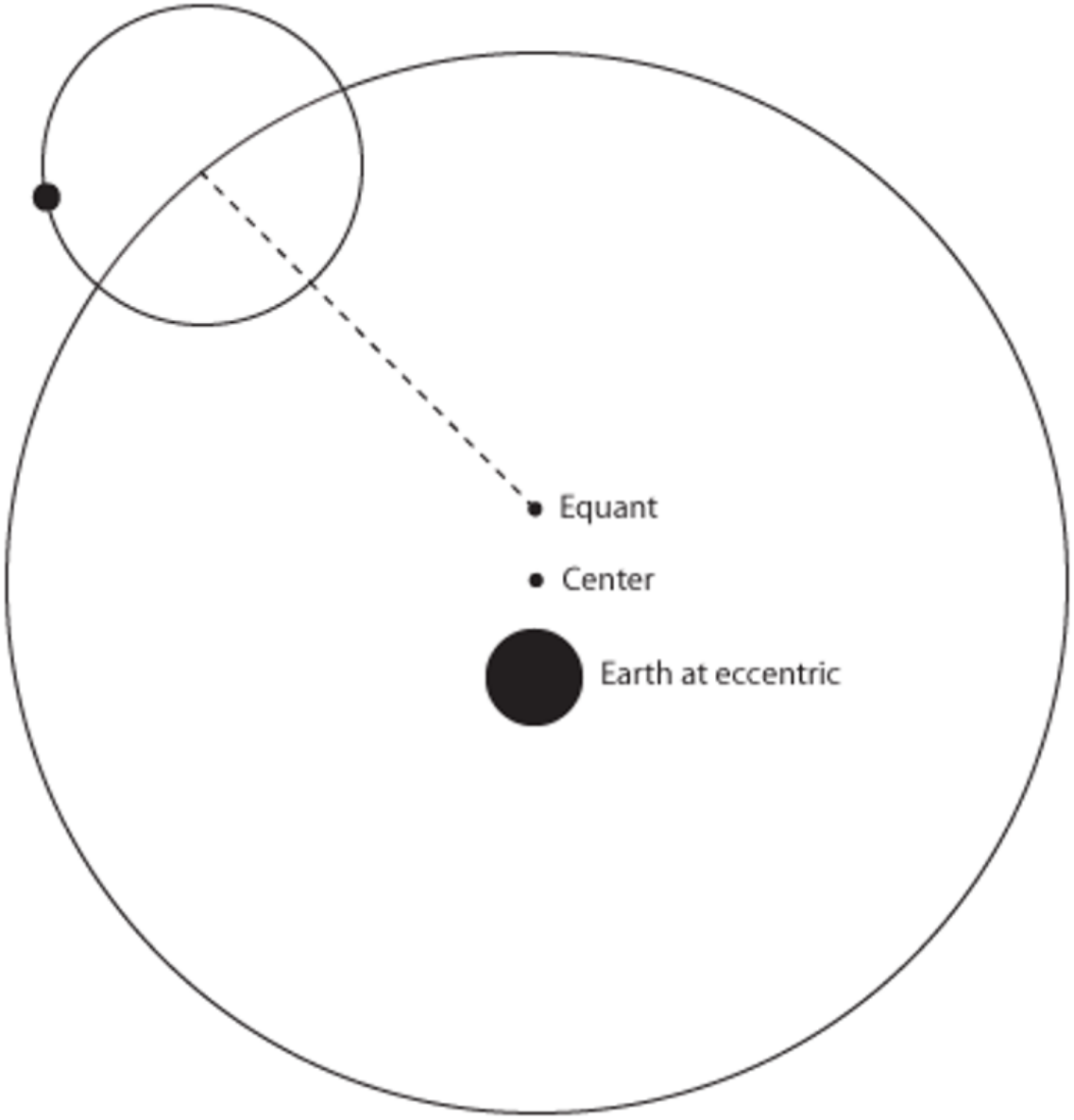
2. Ptolemy

Ptolemy of Alexandria provided the mathematical foundation for geocentrism, around 100 AD. He wanted to explain two problematic observations. First, the planets appear to move at different speeds at different times, contrary to the Aristotelian thesis that they should move with a constant motion. Second, some planets, like Mars, occasionally seem to briefly move backwards in their paths before returning to their regular orbit. Like this:



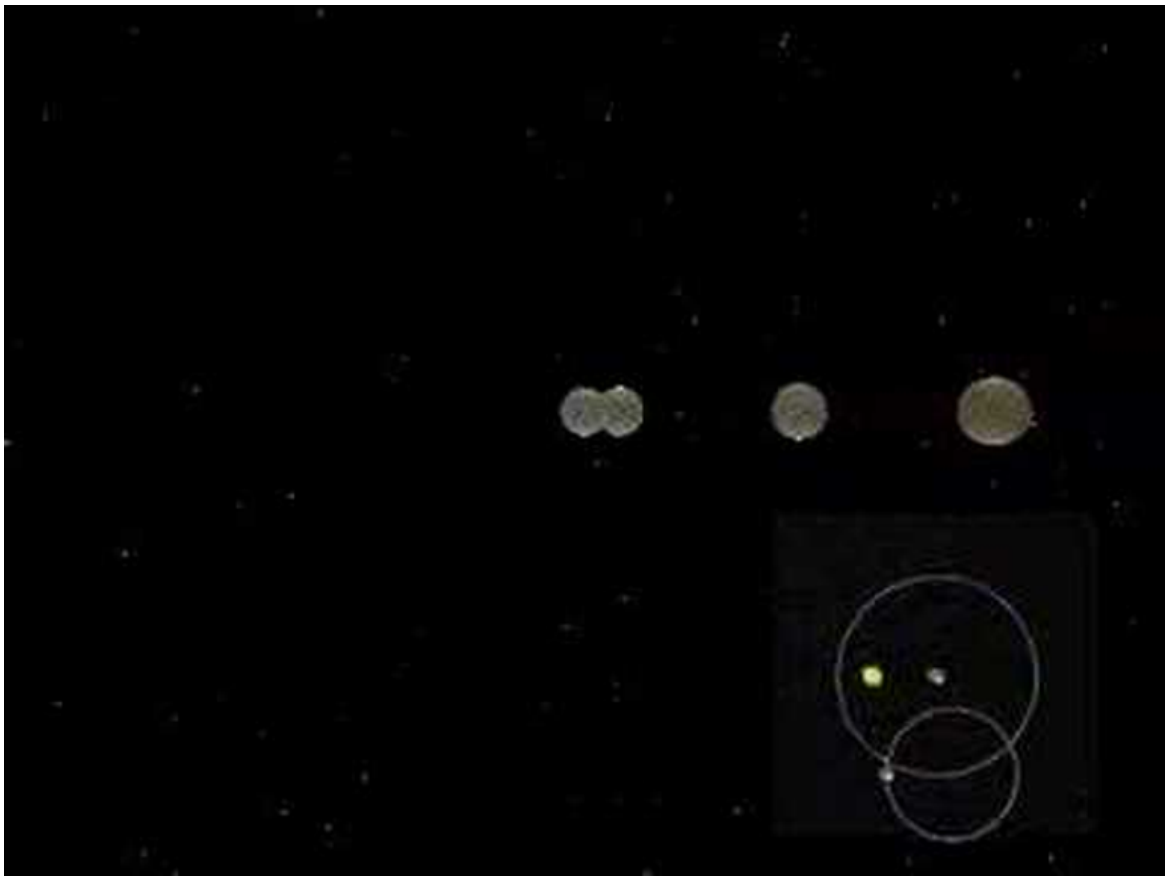
Link to 7-second video.

In order to explain these phenomena, Ptolemy introduced the geometric tools of equants and epicycles. He placed the earth slightly off the center of the planetary orbits, had the planets themselves orbit in little mini-cycles -- so-called “epicycles” -- along their original orbit, and introduced another off-center point, called the equant, in relation to which the motions of the planets are uniform, and which Ptolemy also claimed “controlled” the speed of the planets along their larger orbits. Like this:



[3]

Here’s how these additions make sense of retrograde motion [4]:



Link to 9-second video.

The ability of the Ptolemaic system to account for these phenomena, predicting planetary positions to within a few degrees (Brown, 2016), was a key contributor to its widespread popularity. In fact, the Ptolemaic model is so good that it's still being used to generate celestial motions in planetariums (Wilson, 2000).

3. Copernicus

Copernicus published his heliocentric theory while on his deathbed, in 1543. It retained the circular orbits. More importantly, it of course placed the sun at the centre of the universe and proposed that the earth rotates around its own axis. Copernicus was keen to get rid of Ptolemy's equants, which he abhorred, and instead introduced the notion of an epicyclet (which, to be fair, is kind of just like an equant with its own mini-orbit) [5]. Ptolemy's system had required huge epicycles, and Copernicus was able to substantially reduce their size.

Retrograde motion falls out of his theory like this:

Link to 9-second video.

In order to get the actual motion of the planets correct, both Ptolemy and Copernicus had to bolster their models with many more epicycles, and epicycles upon epicycles, than shown in the above figure and video. Copernicus even considered introducing an epicyclepicyclet -- “an epicyclet whose center was carried round by an epicycle, whose center in turn revolved on the circumference of a deferent concentric with the sun as the center of the universe”... (Complete Dictionary of Scientific Biography, 2008).

Pondering his creation, Copernicus concluded an early manuscript outline his theory thus “Mercury runs on seven circles in all, Venus on five, the earth on three with the moon around it on four, and finally Mars, Jupiter, and Saturn on five each. Thus 34 circles are enough to explain the whole structure of the universe and the entire ballet of the planets” (MacLachlan & Gingerich, 2005).

These inventions might appear like remarkably awkward -- if not ingenious -- ways of making a flawed system fit the observational data. There is however quite an elegant reason why they worked so well: they form a primitive version of Fourier analysis, a modern technique for function approximation. Thus, in the constantly expanding machinery of epicycles and epicyclets, Ptolemy and Copernicus had gotten their hands on a powerful computational tool, which would in fact have allowed them to approximate orbits of a very large number of shapes, including squares and triangles (Hanson, 1960)!

Despite these geometric acrocrabitcs, *Copernicus theory did not fit the available data better than Ptolemy’s*. In the second half of the 16th century, renowned imperial astronomer Tycho Brahe produced the most rigorous astronomical observations to date -- and found that they even fit Copernicus’ data worse than Ptolemy’s in some places (Gingerich, 1973, 1975).

This point seems to have been recognized clearly by enlightenment scholars, many of whom instead chose to praise the increased simplicity and coherence of the Copernican system. However, as just described, it is unclear whether it even offered any such improvements. As Kuhn put it, Copernicus's changes seem "great, yet strangely small", when considering the complexity of the final system (Kuhn, 1957). The mathematician and historian Otto Neugebauer writes:

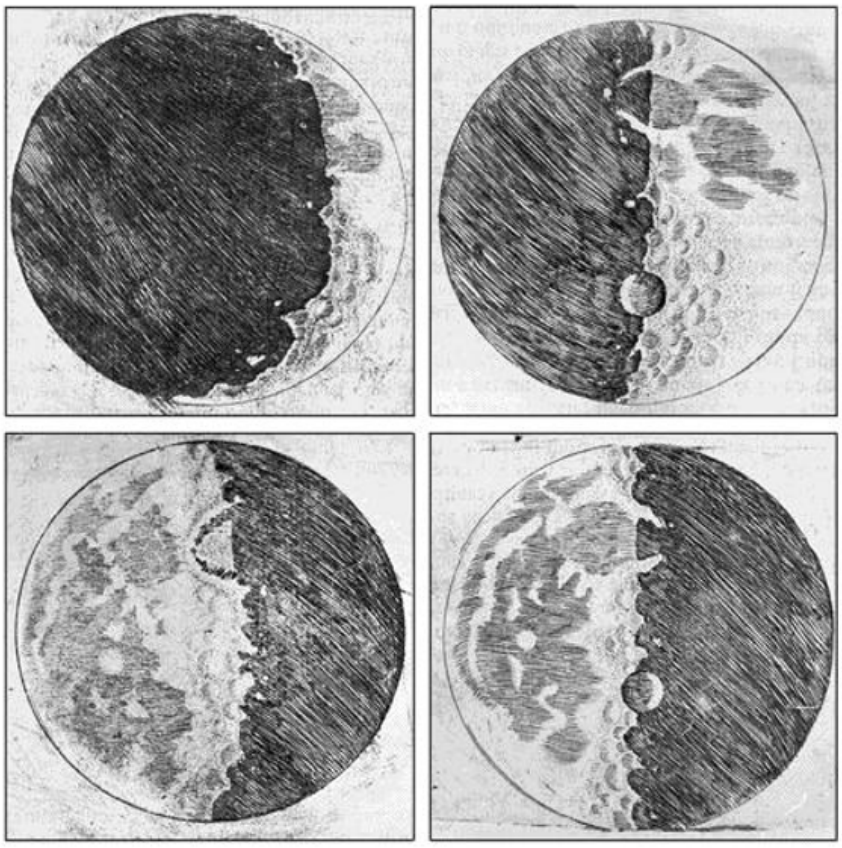
"Modern historians, making ample use of the advantage of hindsight, stress the revolutionary significance of the heliocentric system and the simplifications it had introduced. In fact, the actual computation of planetary positions follows exactly the ancient pattern and the results are the same. [...] Had it not been for Tycho Brahe and Kepler, the Copernican system would have contributed to the perpetuation of the Ptolemaic system in a slightly more complicated form but more pleasing to philosophical minds." (Neugebauer, 1968)

4. Kepler and Galileo

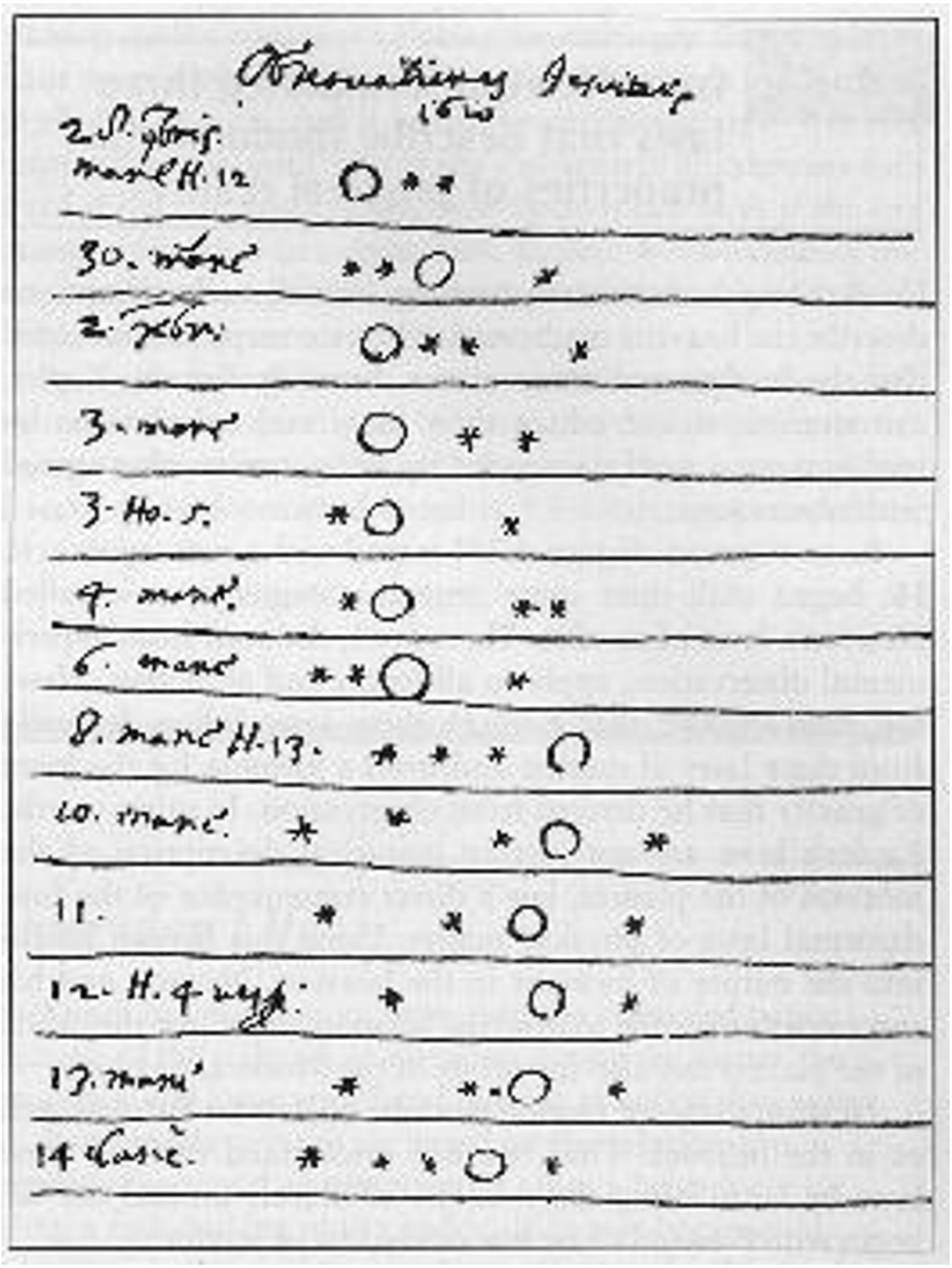
At the turn of the 17th century Kepler, armed with Tycho Brahe's unprecedentedly rigorous data, revised Copernicus' theory and introduced elliptical orbits [6]. He also stopped insisting that the planets follow uniform motions, allowing him to discard the cumbersome epicyclical machinery.

Around the same time Galileo invented the telescope. Upon examining the celestial bodies, he found irregularities that seemed to contradict the Scholastic view of the heavens as a perfect, unchanging realm. There were spots on the sun...

...craters and mountains on the moon...



...and four new moons orbiting Jupiter.



Spurred on by his observations, Galileo would soon begin his ardent defense of heliocentrism. Despite the innovations of Galileo and Kepler, the path ahead wasn’t straightforward.

Galileo focused his arguments on Copernicus' system, *not* Kepler's. And in doing so he faced not only the problems with fitting positional planet data, which Kepler had solved, but also theoretical objections, to which Kepler was still vulnerable.

Consider the tower argument. This is a simple thought experiment: if you drop an object from a tower, it lands right below where you dropped it. But if the earth were moving, shouldn't it instead land some distance away from where you dropped it?

You might feel shocked upon reading the argument, in the same way you might feel shocked by your grandpa making bigoted remarks at the Christmas table, or by a friend trying to recruit you to a pyramid scheme. Just *writing* it, I feel like I'm penning some kind of crackpot, flat-earth polemic. But if the reason is "well obviously it doesn't fall like that… something something Newton…" then remind yourself of the fact that Isaac Newton *had not yet been born*. The dominant physical and cosmological theory of the day was still Aristotle's. If your answer to the tower argument in any way has to invoke Newton, then you likely wouldn't have been able to answer it in 1632.

Did you manage to find some other way of accounting for objects falling down in a straight line from the tower? You might want to take a few minutes to think about it.

[…time for thinking…]

Now if at the end of thinking you convinced yourself of yadda yadda straight line physics yadda yadda you were unfortunately mistaken. The tower argument is *correct*. Objects *do* drift when falling, due to the earth's rotation -- but at a rate which is imperceptible for most plausible tower heights. This is known as the "Coriolis effect", and wasn't properly understood mathematically until the 19th century.

In addition a fair number of astronomical observations seemed to qualitatively contradict heliocentrism -- by leaving out predicted phenomena -- as opposed to just providing quantitative discrepancies in planetary positions. Consider the stellar parallax. A "parallax" is the effect you might have noticed while looking out of a car window, and seeing how things that are closer to you seem to fly by at a faster pace than things farther away. Like this:

Link to 15-second video.

If the earth orbits the sun, something similar should be visible on the night sky, with nearby stars changing their position substantially in relation to more distant stars. Like this:

Link to 22-second video.

No one successfully detected a stellar parallax during the renaissance. This included Tycho, who as mentioned above had gathered the most accurate and exhaustive observations to date. His conclusion was that either the distant stars were so distant that a parallax wasn't detectable using his instruments -- which would entail that space was mostly an unfathomably vast void -- or there simply was no stellar parallax to be detected.

Once again, with the benefit of hindsight it is easy to arbitrate this debate. Space just is really, really, really vast. But it is worth noticing here the similarity to Russell's teapot-style arguments [7]. On two points in a row, the defenders of heliocentrism have been pushed into unfalsifiable territory:

Heliocentrist: "There *is* drift when objects fall from towers -- we just can't measure it!"

Geocentrist: "But provide a phenomenon we *can* measure, then."

Heliocentrist: "Well, according to my recent calculations, a stellar parallax should be observable under these conditions…"

Geocentrist: "But Tycho's data -- the best data astronomical we've ever had -- fails to find any semblance of a parallax. Even Tycho himself thinks the idea is crazy"

Heliocentrist: "The fact that Tycho couldn't detect it doesn't mean it's not there! The stars could be too far away for it to be detected. And things aren't absurd just because prominent scientists say they're absurd"

Geocentrist: "Hold on… not only does your new theory contradict all of established physics, but whenever you're asked for a way to verify it you propose a phenomenon that's barely

testable… and when the tests come out negative you blame the tests and not the theory!”

Heliocentrist: “Okay okay, I’ll give you something else… heliocentrism predicts that Venus will sometimes be on the same side of the sun as Earth, and sometimes on the opposite side…”

Geocentrist: “Yes?”

Heliocentrist: “This means that Venus should appear to vary in size… by...” and the heliocentrist scribbles in his notebook “… as much as… six times.”

And this prediction of the change in size of venus was indeed made by proponents of heliocentrism.

And, once again, although today we know this phenomena does in fact appear, the available observations of the 17th century failed to detect it.

This might all seem messy, complicated, disappointing. If *this* is what the history of intellectual progress actually looks like, how can we ever hope to make deliberate progress in the direction of truth?

It might be helpful to examine a few thinkers -- Copernicus, Kepler, Descartes, Galileo -- who actually accepted heliocentrism, and try to better understand their reasons for doing so.

Little is known about the intellectual development and motivations of Copernicus, as the biography written about him by his sole pupil has been lost. Nonetheless, a tentative suggestion is that he developed rigorous technical knowledge across many fields and found himself in environments which were, if not iconoclastic, at least exceptionally open-minded. According to historian Paul Knoll:

“[The arts faculty at the University of Cracow, where Copernicus studied] held the threefold promise of mathematics and astronomy which were abreast of any developments elsewhere in Europe, of philosophical questioning which undermined much the foundations of much that had been characteristically medieval, and of a critical humanistic attitude which was transforming older cultural and educational values” (Knoll, 1975)

Later, when studying law at the University of Bologna, Copernicus stayed with the astronomy professor Domenico Maria Novara, described as “a mind that dared to challenge the authority of [Ptolemy], the most eminent ancient writer in his chosen fields of study” (Sheila, 2015). Copernicus was also a polymath, who studied law in addition to mathematics and astronomy, and developed an early theory of inflation. His pupil Rheticus was an excellent mathematician, and provided crucial support in helping Copernicus complete his final, major work.

Beyond that some authors claim that Copernicus was influenced by a kind of neoplatonism that regarded the sun as a semi-divine entity, being the source of life and energy -- which made him more content to place it at the centre of the universe (Kuhn, 1957). These claims are however disputed (Sheila, 2015).

These conditions -- technical skill, interdisciplinary knowledge and open-mindedness -- seem necessary for Copernicus development, but they also feel glaringly insufficient.

As for Kepler and Descartes, their acceptance of heliocentrism was not motivated by careful consideration of the available data, but commitments to larger philosophical projects. Kepler is known as a mathematician and astronomer, but in his own day he insisted that he be regarded as a philosopher, concerned with understanding the ultimate nature of the cosmos (Di Liscia, 2017). He did have access to better data -- Tycho’s observations -- than most people before him, and he pored over it with tremendous care. Nonetheless, his preference for elliptical over circular orbits was equally influenced by mystical views regarding the basic

geometric harmony of the universe, in which the sun provided the primary source of motive force (Ladyman, 2001; Di Liscia, 2017; Westman, 2001).

Something similar was true of Descartes, although his underlying philosophical agenda is quite different. A striking example of these commitments is that both Kepler and Descartes argued that a heliocentric world-view was self-evident, in the sense of being derivable from first principles without recourse to empirical observation (Frankfurt, 1999).

Beyond that, I know too little about their respective views to be able to offer any more detailed, mechanistic account of why they preferred heliocentrism.

Galileo -- Copernicus' bulldog -- is a confusing figure as well. Just like Copernicus, Kepler and Descartes, Galileo was not purely guided by careful experiment and analysis of the data -- despite the weight popular history often places upon these characteristics of his. As Einstein writes in his foreword to a modern edition of Galileo's *Dialogue*:

"It has often been maintained that Galileo became the father of modern science by replacing the speculative, deductive method with the empirical, experimental method. I believe, however, that this interpretation would not stand close scrutiny. There is no empirical method without speculative concepts and systems; and there is no speculative thinking whose concepts do not reveal, on closer investigation, the empirical material from which they stem." (Einstein, 2001)

For Galileo, this speculative system consisted in replacing the four Aristotelian elements with a single, unified theory of matter, and replacing the view of nature as a teleological process with a view of it a deterministic, mechanistically intelligible process. Einstein later points out that in some respects this approach was inevitable given the limited experimental methods available to Galileo (for example, he could only measure time intervals longer than a second).

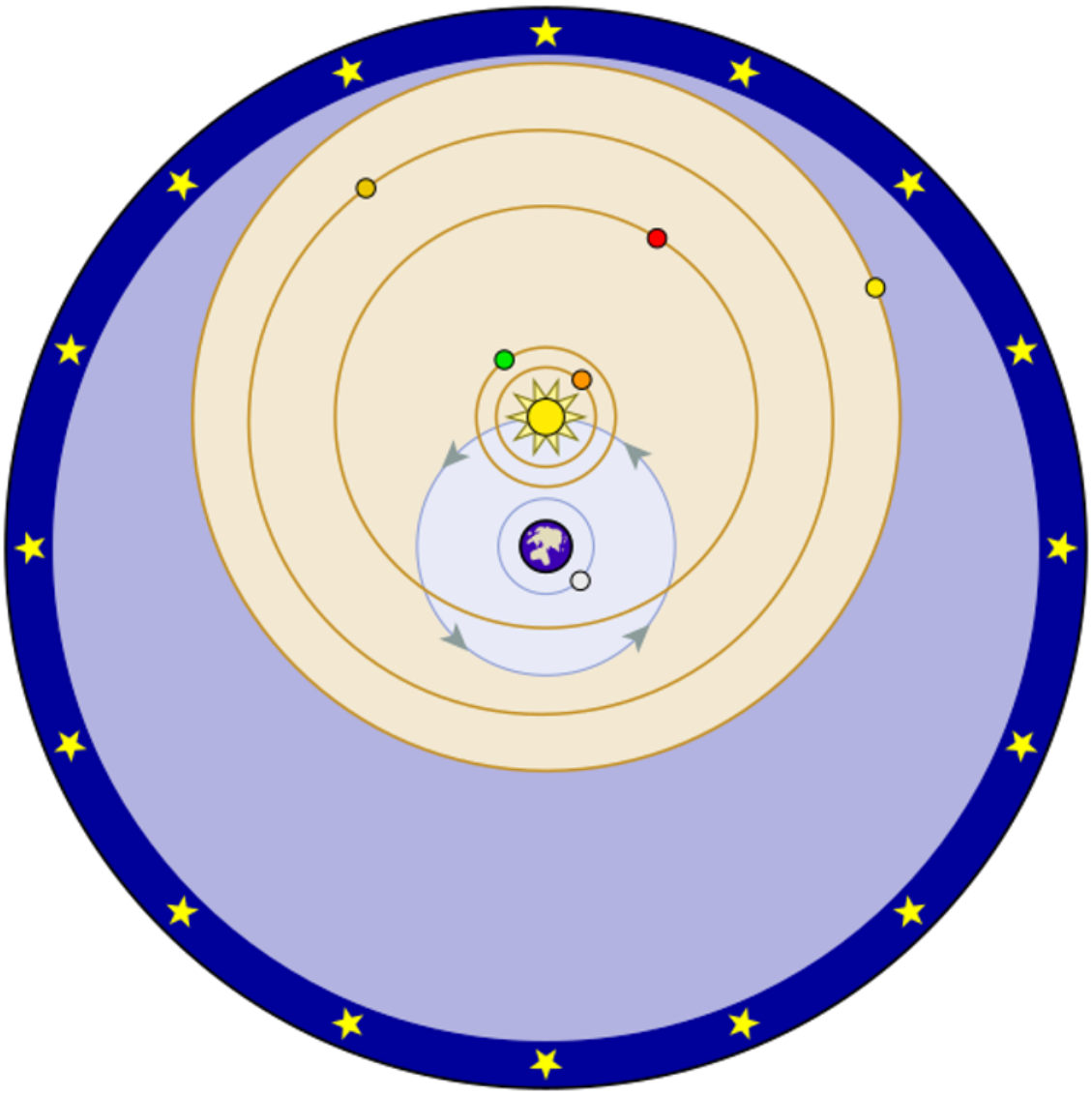
Galileo was also a man of courage and belligerence. One of his strengths was an absolute refusal to accept arguments from authority without experimental evidence or careful reasoning. It appears as if though his belligerence aided him several times in a quite ironic way. Many of the arguments he marshalled against his opponents were either incorrect, or correct but based on incorrect observations. One example is his attempt to derive a theory of tides from the motions of the earth, a project to which he devotes about a fourth of his famous *Dialogue*. Einstein, again, writes "it was Galileo's longing for a mechanical proof of the motion of the earth which misled him into formulating a wrong theory of the tides. [These] fascinating arguments [...] would hardly have been accepted as proofs by Galileo, had his temperament not got the better of him".

Moreover, Galileo's observations of sunspots and moon craters weren't unproblematic. In both cases there is evidence to indicate that he was fooled by optical illusions. And though he *was* also right about the existence of moons orbiting Jupiter, which contradicted the uniqueness of the earth as the only planet with a moon, what he actually observed rather seems to have been Saturn's rings (Ladyman, 2001) [8].

Nonetheless, at this point you might be aching to object that, disregarding inconsistent data, theoretical flaws, failed predictions and incorrect formulation of a theory of the tides... surely Galileo's *Dialogue* provided other convincing arguments that finally tipped the balance in favour of heliocentrism?

Alas, history is *messy*.

Recall that Galileo defended Copernicus system, not Kepler's, and hence had to deal with its flaws. More strikingly, in the above I still haven't mentioned the existence of a third major theory, rivalling both Ptolemy and Copernicus: Tycho Brahe's combined geoheliocentric theory. This theory retained the moon and sun in orbit around the earth but placed all the other planets in orbit around the sun.



Galileo's *Dialogue* does not engage with Tycho's theory *at all*. One suggested explanation (given by an unknown Wikipedia contributor) is that, assuming Galileo's theory of the tides, the Ptolemaic and Tychonic systems are identical, and hence it would suffice to rebut the former. *But the theory of the tides was wrong.*

These theories only differ in their prediction of whether we should be able to observe stellar parallaxes. And as mentioned above, Tycho's data had failed to detect one, which he saw as key evidence for his view.

Eventually though, this historical mess was straightened out, and a crucial experiment arbitrated Galileo, Tycho and Ptolemy. German astronomer Friedrich Bessel's finally managed to observe a stellar parallax *in 1838*. About 200 hundred years later. By *that point*, the Copernican revolution was surely already over -- even the Catholic church had removed Copernicus' *De revolutionibus* from Its index of banned books, as it was simply accepted as true (Lakatos & Zahar, 1975).

5. Newton

At one point Newton also came along, but Galileo died about a year before he was born. Newton's marriage of physics and mathematics, which implied Kepler's laws as a special case, was crucial in demonstrating the viability of heliocentrism. But nonetheless some thinkers did something very right decades before the arrival of the Cambridge genius, which he was very well aware of. For the Copernican revolution might have been completed by Newton, but in the end he still stood on the shoulders of giants.

Now what?

One purpose of this essay has been to portray an important historical era in a more realistic way than other popular portrayals. I asked two questions at the beginning:

1. If you lived in the time of the Copernican revolution, would you have accepted heliocentrism?
2. How should you develop intellectually, in order to become the kind of person who would have accepted heliocentrism during the Copernican revolution?

The preceding section argued that the answer to the first question might quite likely have been no. This section takes a closer look at the second question. I do however want to preface these suggestions by saying that I don't have a good answer to this myself, and suggest you take some time to think of your own answers to these questions. I'd love to hear your thought in the comments.

What about Ibn ash-Shãtir?

There seems to be some Islamic scholars who beat Copernicus to his own game by a few hundred years. I'd be keen to learn more about their story and intellectual habits.

Careful with appearances

Geocentrists liked to claim that it certainly seems like the sun orbits the earth, and not vice versa. There is something odd about this. Consider the following Wittgenstein anecdote:

"He [Wittgenstein] once asked me [Anscombe]: 'Why do people say it is more logical to think that the sun turns around the Earth than Earth rotating around its own axis?' I answered: 'I think because it seems as if the sun turns around the Earth.' 'Good,' he said, 'but how would it have been if it had seemed as if the Earth rotates around its own axis then?'" (Anscombe, 1959)

This quote hopefully inspired in you a lovely sense of confusion. If it it didn't, try reading it again.

When I said above that it certainly seems like the sun orbits the earth and not vice versa, what I meant to say was that it certainly *seems like* it seems like the sun orbits the earth and not vice versa [9].

There's a tendency to use the word "seems" in quite a careless fashion. For example, most people might agree that it seems like, if an astronaut were to push a bowling ball into space, it would eventually slow down and stop, because that's what objects do. At least most people living prior to the 20th century. However, we, *and they*, *already know* that this cannot be true. It suffices to think about the difference between pushing a bowling ball over a carpet, or over a cleaned surface like polished wood, or over ice -- there's a slippery slope here which, if taken to its logical extreme, should make it seem reasonable that a bowling ball wouldn't stop in space. A prompt I find useful is to try to understand why the behaviour of the bowling ball in space could not have been any other way, given how it behaves on earth. That is, trying to understand why, if we genuinely thought a bowling ball would slow down in space, this would entail that the universe was impossibly different from the way it actually is.

Something similar seems true of the feeling that the sun orbits the earth, and this is brought out in the Wittgenstein anecdote. What we think of as “it seems as if though the sun orbits the earth” is actually just us carelessly imposing a mechanism upon a completely different sensation, namely the sensation of “celestial objects seeming to move *exactly as they would move* if the earth orbited the sun and not vice versa”. Whatever it would look like to live in a world where the opposite was true, it certainly wouldn’t look like this.

Careful with your reductios

Many of the major mistakes made by opponents of heliocentrism was to use reductio ad absurdum arguments without really considering whether the conclusion was absurd enough to actually overturn the original argument. Tycho correctly noted that either there wasn’t a stellar parallax or he couldn’t measure it, but incorrectly took the former as more plausible. Proponents of the tower argument *assumed* that objects fall down in straight lines without drift, and that anything else would be percetible by the naked eye. In both cases, people would just have been better off biting the bullet and accepting the implications of heliocentrism. That, of course, raises the question of which bullets one should bite -- and that question is beyond the scope of this essay.

The data is not enough

There’s a naïve view of science according to which the scientist first observes all the available data, then formulates a hypothesis that fits it, and finally tries to falsify this new hypothesis by making a new experiment. The Copernican revolution teaches us that the relation between data and theory is in fact much more subtle than this.

A true theory does not have to immediately explain all the data better than its predecessors, and can remain inconsistent with parts of the data for a long time.

The relation between data and theory is not a one-way shooting range, but an intricate two-way interplay. The data indicates which of our theories are more or less plausible. But our theories also indicate which data is more or less trustworthy [10]. This might seem like a sacrilegious claim to proponents of the naïve view described above: “ignore the data!? That’s just irrational cherry-picking!” Sure, dishonest cherry-picking is bad. Nonetheless, as the Copernican revolution shows, the act of disregarding some data in a principled manner as it doesn’t conform to strong prior expectations has been critical to the progress of science [11].

When Einstein famously remarked “God doesn’t play dice”, he arguably adopted the same kind of mindset. He had built a complex worldview characterised by a certain mathematical law-likeness, and was confident in it to the extent that if quantum mechanics threatened its core principles, then quantum mechanics was wrong -- not him.

Sometimes, scientists have to be bold -- or arrogant -- enough to trust their priors over the data.

...and, finally, deep learning

It seems apt to notice some similarities between the state of astronomy during the Copernican revolution and the current state of deep learning research.

Both are nascent fields, without a unifying theory that can account for the phenomena from first principles, like Newtonian physics eventually did for astronomy.

Both have seen researchers cling to their models for decades without encouraging data: many of the most succesful current deep learning techniques (conv nets, recurrent nets and LSTMs, gradient descent, ...) were invented in the 20th century, but didn’t produce spectacular results until decades later when sufficient computing power became available. It

would be interesting to find out if people like like Geoffrey Hinton and Yann Le Cunn share intellectual habits with people like Copernicus and Galileo.

Finally, I'm particularly struck by the superficial similarities between the way Ptolemy and Copernicus happened upon a general, overpowered tool for function approximation (Fourier analysis) that enabled them to misleadingly gerrymander false theories around the data, and the way modern ML has been criticized as an inscrutable heap of linear algebra and super-efficient GPUs. I haven't explored whether these similarities go any deeper, but one implication seems to be that the power and versatility of deep learning might allow suboptimal architectures to perform deceivingly well (just like the power of epicycle-multiplication kept geocentrism alive) and hence distract us from uncovering the actual architectures underlying cognition and intelligence.

Crossposted to my blog here.

Footnotes

[1] They of course *are* taught, because that is how I learnt about them. But this was in a university course on the philosophy of science. The story of Galileo is probably taught in most middle schools [no source, my own hunch]. But only about 0.5% of US college students major in philosophy [source], and I'd guesstimate something like a third of them to take classes in philosophy of science.

[2] This last step is kind of a blackbox. My model was something like "a true theory was around for long enough, and gained enough support, that it was eventually adopted". This sounds quite romantic, if not magical. It's unclear exactly *how* this happened, and in particular what strategic mistakes the Church made that allowed it to.

[3] Figure credit of the Polaris Institute of Iowa State University, which provides a great tutorial on medieval and renaissance astronomy here.

[4] I spent way too long trying to understand this, but this animation was helpful.

[5] I spent two hours trying to understand the geometry of this and I won't drag you down that rabbit-hole, but if you're keen to explore yourself, check out these links: [1], [2].

[6] It is however a common mistake to imagine these as clearly elongated ellipses: their eccentricity is very small. For most practical purposes apart from measurement and prediction they look like circles (Price, 1957).

[7] Russell's teapot is a skeptic thought-experiment intended to reveal the absurdity of unfalsifiable views, by postulating that there's a teapot orbiting Jupiter and that it's too small to be detectable, but nonetheless insisting that it really is there.

[8] And to think my philosopher friends thinks Gettier problems are nonsense!

[9] There's of course a sense of mysticism in this, which -- like the rest of Wittgenstein's mysticism -- I don't like. Mysticism is mostly just a clever way of scoring social understanding-the-world-points without actually understanding the world. It might be that heliocentrism and geocentrism are genuinely indistinguishable from our vantage point, in which case the confusion here is just a linguistic sleight-of-hand, rather than an actual oddity in how we perceive the world. But this doesn't seem correct. After all, we were able to figure out heliocentrism *from our vantage point*, indicating that heliocentrism is distinguishable from geocentrism from our vantage point.

[10] In Bayesian terms, your posterior is determined by both your likelihoods and your priors.

[11] And is core to rationality itself, on the Bayesian view.

References

Anscombe, E. (1959). *An Introduction to Wittgenstein's Tractatus.* pp. 151.

Brown, M. (2016) "Copernicus' revolution and Galileo's vision: our changing view of the universe in pictures". *The Conversation*. Available online here.

"Copernicus, Nicholas." Complete Dictionary of Scientific Biography. Retrieved October 26, 2017 from Encyclopedia.com, here.

Di Liscia, D. A. "Johannes Kepler". *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Zalta, E. N. (ed.).

Einstein, A. (2001). (Foreword) *Dialogue Concerning the Two Chief World Systems*: Ptolemaic and Copernican.

Frankfurt, H. (1999), *Necessity, Volition and Love*. pp. 40.

Gingerich, O. J. (1973) "The Copernican Celebration". *Science Year*, pp. 266-267.

Gingerich, O. J. (1975) "'Crisis' versus aesthetic in the Copernican revolution". *Vistas in Astronomy* 17(1), pp. 85-95.

Hanson, N. R. (1960) "The Mathematical Power of Epicyclical Astronomy" *Isis*, 51(2), pp. 150-158.

Knoll, P. (1975) "The Arts Faculty at the University of Cracow at the end of the Fifteenth Century". *The Copernican Achievement*, Westman, R. S (ed.)

Kuhn, T. (1957) *The Copernican Revolution*. pp. 133.

Ladyman, J. (2001) *Understanding Philosophy of Science*. Chapter 4: Revolutions and Rationality.

Lakatos, I. & Zahar, E. (1975). "Why did Copernicus Research Program Supersede Ptolemy's?". *The Copernican Achievement*, Westman, R. S (ed.)

MacLachlan, J & Gingerich, O. J. (2005) *Nicolaus Copernicus: Making the Earth a Planet*, pp. 76.

Neugebauer, O. (1968). "On the Planetary Theory of Copernicus", *Vistas in Astronomy.*,10, pp. 103.

Sheila, R. "Nicolaus Copernicus". *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), Zalta, E. N. (ed.), available online.

Price, D. J. (1957) "Contra Copernicus: a critical re-estimation of the mathematical Planetary Theory of Copernicus, Ptolemy and Kepler". *Critical Problems in the History of Science*, Clagett, M. (ed.).

Westman, R. S. (2001) "Kepler's early physical-astrological problematic." *Journal for the History of Astronomy,* 32, pp. 227-236.

Wilson, L. A. (2000) "The Ptolemaic Model" in the Polaris Project, Iowa State University. Available online here.

Announcing the AI Alignment Prize

Stronger than human artificial intelligence would be dangerous to humanity. It is vital any such intelligence's goals are aligned with humanity's goals. Maximizing the chance that this happens is a difficult, important and under-studied problem.

To encourage more and better work on this important problem, we (Zvi Mowshowitz and Vladimir Slepnev) are announcing a $5000 prize for publicly posted work advancing understanding of AI alignment, funded by Paul Christiano.

This prize will be awarded based on entries gathered over the next two months. If the prize is successful, we will award further prizes in the future.

The prize is not backed by or affiliated with any organization.

Rules

Your entry must be published online for the first time between November 3 and December 31, 2017, and contain novel ideas about AI alignment. Entries have no minimum or maximum size. Important ideas can be short!

Your entry must be written by you, and submitted before 9pm Pacific Time on December 31, 2017. Submit your entries either as links in the comments to this post, or by email to apply@ai-alignment.com. We may provide feedback on early entries to allow improvement.

We will award $5000 to between one and five winners. The first place winner will get at least $2500. The second place winner will get at least $1000. Other winners will get at least $500.

Entries will be judged subjectively. Final judgment will be by Paul Christiano. Prizes will be awarded on or before January 15, 2018.

What kind of work are we looking for?

AI Alignment focuses on ways to ensure that future smarter than human intelligence will have goals aligned with the goals of humanity. Many approaches to AI Alignment deserve attention. This includes technical and philosophical topics, as well as strategic research about related social, economic or political issues. A non-exhaustive list of technical and other topics can be found here.

We are **not** interested in research dealing with the dangers of existing machine learning systems commonly called AI that do not have smarter than human intelligence. These concerns are also understudied, but are not the subject of this prize except in the context of future smarter than human intelligence. We are also **not** interested in general AI research. We care about AI alignment, which may or may not also advance the cause of general AI research.

(Addendum: the results of the prize and the rules for the next round have now been announced.)

The Archipelago Model of Community Standards

Epistemic Status: My best guess. I don't know if this will work but it seems like the obvious experiment to try more of.

Epistemic Effort: Spent several months thinking casually, 25ish minutes consolidating earlier memories and concerns, and maybe 10ish minutes thinking about potential predictions. See comment.

Building off:

- Open Problems in Group Rationality [Conor Moreton]
- Archipelago and Atomic Commutarianism [Scott Alexander]

Claim 1 - If you are dissatisfied with the norms/standards in a vaguely defined community, a good first step is to **refactor that community into sub-groups with clearly defined goals and leadership.**

Claim 2 - People have different goals, and you may be wrong about what norms are important even given a certain goal. So, also consider ***proactively cooperating*** **with other people forming alternate subgroups** out of the same parent group, with the goal of learning from each other.

Refactoring Into Subcommunities

Building groups that accomplish anything is hard. Building groups that prioritize independent thinking to solve novel problems is harder. But when faced with a hard problem, a useful technique is to refactor it into something simpler.

In "Open Problems in Group Rationality", Conor lists several common tensions. I include them here for reference (although *any* combination of difficult group rationality problems would suffice to motivate this post).

1. Buy-in and retention.
2. Defection and discontent.
3. Safety versus standards.
4. Productivity versus relevance.
5. Sovereignty versus cooperation.
6. Moloch and the problem of distributed moral action.

These problems *don't go away* when you have clearly defined goals. A corporation with a clearcut mission and strategy (i.e maximize profit by selling widgets) still has to navigate the balance of "hold their employees to a high standards to increase performance" and "make sure employees feel safe enough to do good work without getting wracked with anxiety" (or, just quit).

Such a corporation might make different tradeoffs in different situations - if there's a labor surplus, they might be less worried about employees quitting because they can just find more. If the job involves creative knowledge work, anxiety might have greater costs to productivity. Or maybe they're *not* just profit-maximizing: maybe the CEO cares about employee mental health for its own sake.

But well defined goals, with leaders who can enforce them, at least makes it *possible* to figure out what tradeoffs to make and actually make them.

Whereas if you live in a loosely defined community where people show up and leave whenever they want, *and nobody can even precisely agree on what the community is,* you'll have a lot more trouble.

People who care a lot about, say, personal sovereighty, will constantly push for norms that maximize freedom. People that care about cooperation will push for norms encouraging everyone to work harder and be more reliabl at personal freedom's expense.

Maybe one group can win - possibly by persuading everyone they are right, or simply by being more numerous.

But,

A) You probably can't win every cultural battle.

B) Even if you could, you'd spend a lot of time and energy fighting that might be better spent actually accomplishing whatever these norms are actually *for.*

So if you can manage to *avoid* infighting while still accomplishing your goals, all things being equal that's preferable.

Considering Archipelago

Once this thought occured to me, I was immediately reminded of Scott Alexander's Archipelago concept. A quick recap:

Imagine a bunch of factions fighting for political control over a country. They've agreed upon the strict principle of harm (no physically hurting or stealing from each other). But they still disagree on things like "does pornography harm people", "do cigarette ads harm people", "does homosexuality harm the institution of marriage which in turn harms people?", "does soda harm people", etc.

And this is bad not just because everyone wastes all this time fighting over norms, but because the nature of their disagreement incentivizes them to fight over *what harm even **is**.*

And this in turn incentivizes them to fight over both definitions of words (distracting and time-wasting) and *what counts as evidence or good reasoning* through a politically motivated lens. (Which makes it harder to *ever* use evidence and reasoning to resolve issues, even uncontroversial ones)

Then...

Imagine someone discovers an archipelago of empty islands. And instead of continuing to fight, the people who want to live in Sciencetopia go off to found an island-state based on ideal scientific processes, and the people who want to live in Libertopia go off and found a society based on the strict principle of harm, and the people who want to live in Christiantopia go found a fundamentalist Christian commune.

They agree on an overarching set of rules, paying some taxes to a central authority that handles things like "dumping pollutants into the oceans/air that would affect other islands" and "making sure children are well educated enough to have the opportunity to understand why they might consider moving to other islands."

Practical Applications

There's a bunch of reasons the Archipelago concept doesn't work as well in practice. There are no magical empty islands we can just take over. Leaving a place if you're unhappy is harder than it sounds. Resolving the "think of the children" issue will be very contentious.

But, we don't need perfect-idealized-archipelago to make use of the general concept. We don't even need a broad critical mass of change.

You, personally, could just do something with it, right now.

If you have an event you're running, or an online space that you control, or an organization you run, you can set the norms. Rather than opting-by-default into the generic average norms of your peers, you can say "This is a space specifically for X. If you want to participate, you will need to hold yourself to Y particular standard."

Some features and considerations:

You Can Test More Interesting Ideas. If a hundred people have to agree on something, you'll only get to try things that you can can 50+ people on board with (due to crowd inertia, regardless of whether you have a formal democracy)

But maybe you can get 10 people to try a more extreme experiment. (And if you share knowledge, both about experiments that work and ones that don't, you can build the overall body of community-knowledge in your social world)

I would rather have a world where 100 people try 10 different experiments, even if I *disagree* with most of those experiments and wouldn't want to participate myself.

You Can Simplify the Problem and Isolate Experimental Variables. "Good" science tests a single variable at the time so you can learn more about what-causes-what.

In practice, if you're building an organization, you may not have time to do "proper science" - you may need to get a group working ASAP, and you may need to test a few ideas at once to have a chance at success.

But, all things being equal it's still convenient to isolate factors as much as possible. One benefit to refactoring a community into smaller pieces is you can pick more specific goals. Instead of reinventing every single wheel at once, pick a few specific axes you're trying to learn about.

This will both make the problem easier, as well as make it easier to *learn from.*

You Can 'Timeshare Islands'. Maybe you don't have an entire space that you can control. But maybe you and some other friends have a shared space. (Say, a weekly meetup).

Instead of having the meetup be a generic thing catering to the average common denominator of members, you can collectively agree to use it for experiments (at least sometimes). Make it easier for one person to say 'Okay, this week I'd like to run an activity that'll require different norms than we're used to. Please come prepared for things to be a bit different.'

This comes with some complications - one of the benefits of a recurring event is people roughly know what to expect, so it may not be good to do this all the time. But generally, giving the person running a given event the authority to try some different norms out can get you some of the benefits of the Archipelago concept.

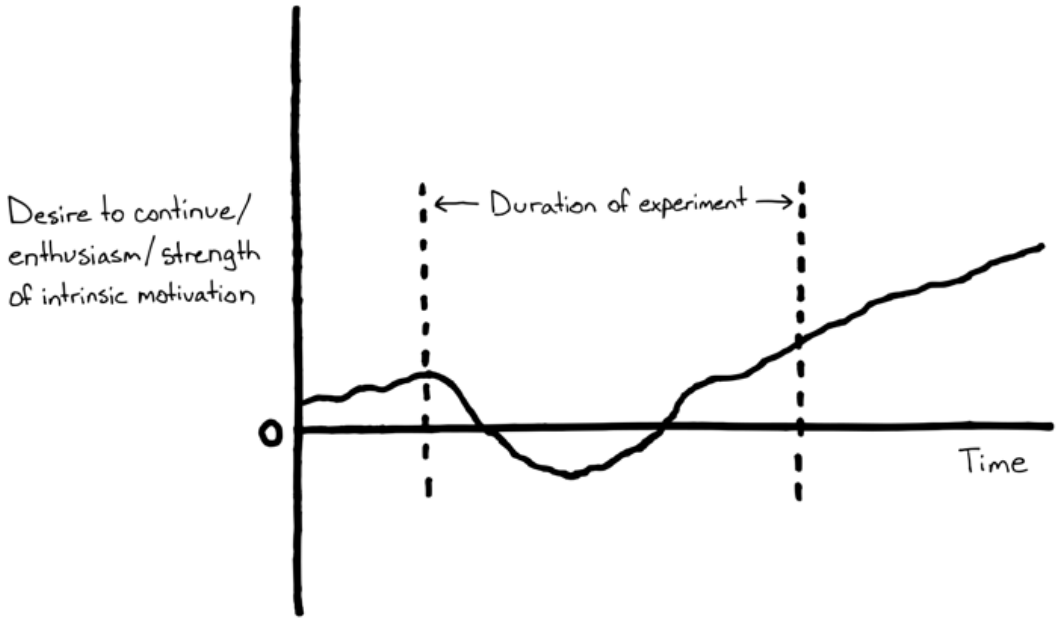
You Can Start With Just *One* Meetup

Viliam in the comments made a note I wanted to include here:

> It is important to notice that the "island" doesn't have to be fully built from start. "Let's start a new subgroup" sounds scary; too much responsibility and possibly not enough status. "Let's have *one meeting* where we try the norm X and see how it works" sounds much easier; and if it works, people would be more willing to have another meeting like that, possibly leading to the creation of a new community.

Making It Through the 'Unpleasant Valley' of *Group* Experimentation.

I think this graph was underappreciated in its original post. When people try new things (a new diet or exercise program, studying a new skill, etc), the new thing involves effort and challenges that in some ways make it seem worse than whatever their default behavior was.



Some experiments are just duds. But oftentimes it *feels* like it'll turn out to be a dud, when you're in the Unpleasant Valley, and in fact you just haven't stuck with it long enough for it to bear fruit.

This is hard enough for *solo* experiments. For group experiments, where not just one but *many* people must all try a thing at once and *get good at it*, all it takes is a little defection to spiral into a mass exodus.

Refactoring communities into smaller groups with clear subgoals can make it *possible* for a group to make it through the Valley of Unpleasantness together.

Overlapping Social Spheres

Sharing Islands and Cross Pollination

In the end, I don't think "Islands" is quite the right metaphor here. One of the things that makes *social* archipelago different from the canonical example is that the islands overlap. People may be a member of multiple groups and sub-groups.

A benefit of this is cross pollination - it's easier to share information and grow if you have people who exist in multiple subcultures (sub-subcultures?) and can translate ideas between them.

How much benefit this yields depends on how mindfully people are approaching the concept, and how much of their ideas they are sharing (making both the object-level-idea and the underlying reasons accessible to others).

This post is primarily intended as reference - I have more specific ideas on what kinds of communities *I* want to participate in, and thoughts on "underexplored social niches" that I think others might consider experimenting with. Some of those thoughts will be on the LessWrong front page, others on my private profile or the Meta section.

But meanwhile, I hope to see more groups of people in my filter bubble self organizing, carving out spaces to try novel concepts.

Mosquito killing begins

This is a linkpost for https://qz.com/1123456/the-epa-has-approved-the-release-of-weaponized-mosquitoes-in-20-us-states/

Security Mindset and Ordinary Paranoia

Follow-up to: AI Alignment: Why It's Hard, and Where to Start

---

(AMBER, *a philanthropist interested in a more reliable Internet, and* CORAL, *a computer security professional, are at a conference hotel together discussing what Coral insists is a difficult and important issue: the difficulty of building "secure" software.*)

AMBER: So, Coral, I understand that you believe it is very important, when creating software, to make that software be what you call "secure".

CORAL: Especially if it's connected to the Internet, or if it controls money or other valuables. But yes, that's right.

AMBER: I find it hard to believe that this needs to be a separate topic in computer science. In general, programmers need to figure out how to make computers do what they want. The people building operating systems surely won't want them to give access to unauthorized users, just like they won't want those computers to crash. Why is one problem so much more difficult than the other?

CORAL: That's a deep question, but to give a partial deep answer: When you expose a device to the Internet, you're potentially exposing it to intelligent adversaries who can find special, weird interactions with the system that make the pieces behave in weird ways that the programmers did not think of. When you're dealing with that kind of problem, you'll use a different set of methods and tools.

AMBER: Any system that crashes is behaving in a way the programmer didn't expect, and programmers already need to stop that from happening. How is this case different?

CORAL: Okay, so... imagine that your system is going to take in one kilobyte of input per session. (Although that itself is the sort of assumption we'd question and ask what happens if it gets a megabyte of input instead—but never mind.) If the input is one kilobyte, then there are $2^{8,000}$ possible inputs, or about $10^{2,400}$ or so. Again, for the sake of extending the simple visualization, imagine that a computer gets a billion inputs per second. Suppose that only a googol, 10^{100}, out of the $10^{2,400}$ possible inputs, cause the system to behave a certain way the original designer didn't intend.

If the system is getting inputs in a way that's uncorrelated with whether the input is a misbehaving one, it won't hit on a misbehaving state before the end of the universe. If there's an intelligent adversary who understands the system, on the other hand, they may be able to find one of the very rare inputs that makes the system misbehave. So a piece of the system that would literally never in a million years misbehave on random inputs, may break when an intelligent adversary tries deliberately to break it.

AMBER: So you're saying that it's more difficult because the programmer is pitting their wits against an adversary who may be more intelligent than themselves.

CORAL: That's an almost-right way of putting it. What matters isn't so much the "adversary" part as the optimization part. There are systematic, nonrandom forces strongly selecting for particular outcomes, causing pieces of the system to go down weird execution paths and occupy unexpected states. If your system literally has no misbehavior modes at all, it doesn't matter if you have IQ 140 and the enemy has IQ 160—it's not an arm-wrestling contest. It's just very much harder to build a system that doesn't enter weird states when the weird states are being selected-for in a correlated way, rather than happening only by accident. The weirdness-selecting forces can search through parts of the larger state space that you yourself failed to imagine. Beating that does indeed require new skills and a different mode of thinking, what Bruce Schneier called "security mindset".

AMBER: Ah, and what is this security mindset?

CORAL: I can say one or two things about it, but keep in mind we are dealing with a quality of thinking that is not entirely effable. If I could give you a handful of platitudes about security mindset, and that would actually cause you to be able to design secure software, the Internet would look very different from how it presently does. That said, it seems to me that what has been called "security mindset" can be divided into two components, one of which is much less difficult than the other. And this can fool people into overestimating their own safety, because they can get the easier half of security mindset and overlook the other half. The less difficult component, I will call by the term "ordinary paranoia".

AMBER: *Ordinary* paranoia?

CORAL: Lots of programmers have the ability to imagine adversaries trying to threaten them. They imagine how likely it is that the adversaries are able to attack them a particular way, and then they try to block off the adversaries from threatening that way. Imagining attacks, including weird or clever attacks, and parrying them with measures you imagine will stop the attack; that is ordinary paranoia.

AMBER: Isn't that what security is all about? What do you claim is the other half?

CORAL: To put it as a platitude, I might say… defending against mistakes in your own assumptions rather than against external adversaries.

AMBER: Can you give me an example of a difference?

CORAL: An ordinary paranoid programmer imagines that an adversary might try to read the file containing all the usernames and passwords. They might try to store the file in a special, secure area of the disk or a special subpart of the operating system that's supposed to be harder to read. Conversely, somebody with security mindset thinks, "No matter what kind of special system I put around this file, I'm disturbed by needing to make the assumption that this file can't be read. Maybe the special code I write, because it's used less often, is more likely to contain bugs. Or maybe there's a way to fish data out of the disk that doesn't go through the code I wrote."

AMBER: And they imagine more and more ways that the adversary might be able to get at the information, and block those avenues off too! Because they have better imaginations.

CORAL: Well, we kind of do, but that's not the key difference. What we'll really want to do is come up with a way for the computer to check passwords that doesn't rely on the computer storing the password *at all, anywhere*.

AMBER: Ah, like encrypting the password file!

CORAL: No, that just duplicates the problem at one remove. If the computer can decrypt the password file to check it, it's stored the decryption key somewhere, and the attacker may be able to steal that key too.

AMBER: But then the attacker has to steal two things instead of one; doesn't that make the system more secure? Especially if you write two different sections of special filesystem code for hiding the encryption key and hiding the encrypted password file?

CORAL: That's exactly what I mean by distinguishing “ordinary paranoia” that doesn't capture the full security mindset. So long as the system is capable of reconstructing the password, we'll always worry that the adversary might be able to trick the system into doing just that. What somebody with security mindset will recognize as a deeper solution is to store a one-way hash of the password, rather than storing the plaintext password. Then even if the attacker reads off the password file, they still can't give what the system will recognize as a password.

AMBER: Ah, that's quite clever! But I don't see what's so qualitatively different between that measure, and my measure for hiding the key and the encrypted password file separately. I agree that your measure is more clever and elegant, but of course you'll know better standard solutions than I do, since you work in this area professionally. I don't see the qualitative line dividing your solution from my solution.

CORAL: Um, it's hard to say this without offending some people, but... it's possible that even after I try to explain the difference, which I'm about to do, you won't get it. Like I said, if I could give you some handy platitudes and transform you into somebody capable of doing truly good work in computer security, the Internet would look very different from its present form. I can try to describe one aspect of the difference, but that may put me in the position of a mathematician trying to explain what looks more promising about one proof avenue than another; you can listen to everything they say and nod along and still not be transformed into a mathematician. So I *am* going to try to explain the difference, but again, I don’t know of any simple instruction manuals for becoming Bruce Schneier.

AMBER: I confess to feeling slightly skeptical at this supposedly ineffable ability that some people possess and others don't—

CORAL: There are things like that in many professions. Some people pick up programming at age five by glancing through a page of BASIC programs written for a TRS-80, and some people struggle really hard to grasp basic Python at age twenty-five. That's not because there's some mysterious truth the five-year-old knows that you can verbally transmit to the twenty-five-year-old.

And, yes, the five-year-old will become far better with practice; it's not like we're talking about untrainable genius. And there may be platitudes you can tell the 25-year-old that will help them struggle a little less. But sometimes a profession requires thinking in an unusual way and some people's minds more easily turn sideways in that particular dimension.

AMBER: Fine, go on.

CORAL: Okay, so... you thought of putting the encrypted password file in one special place in the filesystem, and the key in another special place. Why not encrypt the key too, write a third special section of code, and store the key to the encrypted key there? Wouldn't that make the system even more secure? How about seven keys hidden in different places, wouldn't that be extremely secure? Practically unbreakable, even?

AMBER: Well, that version of the idea does feel a little silly. If you're trying to secure a door, a lock that takes two keys might be more secure than a lock that only needs one key, but seven keys doesn't feel like it makes the door that much more secure than two.

CORAL: Why not?

AMBER: It just seems silly. You'd probably have a better way of saying it than I would.

CORAL: Well, a fancy way of describing the silliness is that the chance of obtaining the seventh key is not conditionally independent of the chance of obtaining the first two keys. If I can read the encrypted password file, and read your encrypted encryption key, then I've probably come up with something that just bypasses your filesystem and reads directly from the disk. And the more complicated you make your filesystem, the more likely it is that I can find a weird system state that will let me do just that. Maybe the special section of filesystem code you wrote to hide your fourth key is the one with the bug that lets me read the disk directly.

AMBER: So the difference is that the person with a *true* security mindset found a defense that makes the system simpler rather than more complicated.

CORAL: Again, that's almost right. By hashing the passwords, the security professional has made their *reasoning* about the system less complicated. They've eliminated the need for an assumption that might be put under a lot of pressure. If you put the key in one special place and the encrypted password file in another special place, the system as a whole is still able to decrypt the user's password. An adversary probing the state space might be able to trigger that password-decrypting state because the system is designed to do that on at least some occasions. By hashing the password file we eliminate that whole internal debate from the reasoning on which the system's security rests.

AMBER: But even after you've come up with that clever trick, something could still go wrong. You're still not absolutely secure. What if somebody uses “password” as their password?

CORAL: Or what if somebody comes up a way to read off the password after the user has entered it and while it's still stored in RAM, because something got access to RAM? The point of eliminating the extra assumption from the reasoning about the system's security is not that we are then absolutely secure and safe and can relax. Somebody with security mindset is *never* going to be that relaxed about the edifice of reasoning saying the system is secure.

For that matter, while there are some normal programmers doing normal programming who might put in a bunch of debugging effort and then feel satisfied, like they'd done all they could reasonably do, programmers with decent levels of ordinary paranoia about ordinary programs will go on chewing ideas in the shower and coming up with more function tests for the system to pass. So the distinction between security mindset and ordinary paranoia isn't that ordinary paranoids will relax. It's

that... again to put it as a platitude, the ordinary paranoid is running around putting out fires in the form of ways they imagine an adversary might attack, and somebody with security mindset is defending against something closer to “what if an element of this reasoning is mistaken”. Instead of trying really hard to ensure nobody can read a disk, we are going to build a system that's secure even if somebody does read the disk, and *that* is our first line of defense. And then we are also going to build a filesystem that doesn't let adversaries read the password file, as a *second* line of defense in case our one-way hash is secretly broken, and because there's no positive need to let adversaries read the disk so why let them. And then we're going to salt the hash in case somebody snuck a low-entropy password through our system and the adversary manages to read the password anyway.

AMBER: So rather than trying to outwit adversaries, somebody with true security mindset tries to make fewer assumptions.

CORAL: Well, we think in terms of adversaries too! Adversarial reasoning is easier to teach than security mindset, but it's still (a) mandatory and (b) hard to teach in an absolute sense. A lot of people can't master it, which is why a description of “security mindset” often opens with a story about somebody failing at adversarial reasoning and somebody else launching a clever attack to penetrate their defense.

You need to master two ways of thinking, and there are a lot of people going around who have the first way of thinking but not the second. One way I'd describe the deeper skill is seeing a system's security as resting on a story about why that system is safe. We want that safety-story to be as solid as possible. One of the implications is resting the story on as few assumptions as possible; as the saying goes, the only gear that never fails is one that has been designed out of the machine.

AMBER: But can't you also get better security by adding more lines of defense? Wouldn't that be more complexity in the story, and also better security?

CORAL: There's also something to be said for preferring disjunctive reasoning over conjunctive reasoning in the safety-story. But it's important to realize that you do want a primary line of defense that is supposed to just work and be unassailable, not a series of weaker fences that you think might maybe work. Somebody who doesn't understand cryptography might devise twenty clever-seeming amateur codes and apply them all in sequence, thinking that, even if one of the codes turns out to be breakable, surely they won't *all* be breakable. The NSA will assign that mighty edifice of amateur encryption to an intern, and the intern will crack it in an afternoon. There's something to be said for redundancy, and having fallbacks in case the unassailable wall falls; it can be wise to have additional lines of defense, so long as the added complexity does not make the larger system harder to understand or increase its vulnerable surfaces. But at the core you need a simple, solid story about why the system is secure, and a good security thinker will be trying to eliminate whole assumptions from that story and strengthening its core pillars, not only scurrying around parrying expected attacks and putting out risk-fires.

That said, it's better to use two true assumptions than one false assumption, so simplicity isn't everything.

AMBER: I wonder if that way of thinking has applications beyond computer security?

CORAL: I'd rather think so, as the proverb about gears suggests.

For example, stepping out of character for a moment, the author of this dialogue has sometimes been known to discuss the alignment problem for Artificial General Intelligence. He was talking at one point about trying to measure rates of improvement inside a growing AI system, so that it would not do too much thinking with humans out of the loop if a breakthrough occurred while the system was running overnight. The person he was talking to replied that, to him, it seemed unlikely that an AGI would gain in power that fast. To which the author replied, more or less:

> It shouldn't be your job to guess how fast the AGI might improve! If you write a system that will hurt you *if* a certain speed of self-improvement turns out to be possible, then you've written the wrong code. The code should just never hurt you regardless of the true value of that background parameter.
>
> A better way to set up the AGI would be to measure how much improvement is taking place, and if more than *X* improvement takes place, suspend the system until a programmer validates the progress that's already occurred. That way even if the improvement takes place over the course of a millisecond, you're still fine, so long as the system works as intended. Maybe the system doesn't work as intended because of some other mistake, but that's a better problem to worry about than a system that hurts you *even if* it works as intended.
>
> Similarly, you want to design the system so that if it discovers amazing new capabilities, it waits for an operator to validate use of those capabilities—not rely on the operator to watch what's happening and press a suspend button. You shouldn't rely on the speed of discovery or the speed of disaster being less than the operator's reaction time. There's no *need* to bake in an assumption like that if you can find a design that's safe regardless. For example, by operating on a paradigm of allowing operator-whitelisted methods rather than avoiding operator-blacklisted methods; you require the operator to say "Yes" before proceeding, rather than assuming they're present and attentive and can say "No" fast enough.

AMBER: Well, okay, but if we're guarding against an AI system discovering cosmic powers in a millisecond, that does seem to me like an unreasonable thing to worry about. I guess that marks me as a merely ordinary paranoid.

CORAL: Indeed, one of the hallmarks of security professionals is that they spend a lot of time worrying about edge cases that would fail to alarm an ordinary paranoid because the edge case doesn't sound like something an adversary is likely to do. Here's an example from the Freedom to Tinker blog:

> This interest in "harmless failures" – cases where an adversary can cause an anomalous but not directly harmful outcome – is another hallmark of the security mindset. Not all "harmless failures" lead to big trouble, but it's surprising how often a clever adversary can pile up a stack of seemingly harmless failures into a dangerous tower of trouble. Harmless failures are bad hygiene. We try to stamp them out when we can...
>
> To see why, consider the donotreply.com email story that hit the press recently. When companies send out commercial email (e.g., an airline notifying a passenger of a flight delay) and they don't want the recipient to reply to the email, they often put in a bogus From address like donotreply@donotreply.com. A clever guy registered the domain donotreply.com, thereby receiving all email addressed to donotreply.com. This included "bounce" replies to misaddressed emails, some of

> which contained copies of the original email, with information such as bank account statements, site information about military bases in Iraq, and so on...
>
> The people who put donotreply.com email addresses into their outgoing email must have known that they didn't control the donotreply.com domain, so they must have thought of any reply messages directed there as harmless failures. Having gotten that far, there are two ways to avoid trouble. The first way is to think carefully about the traffic that might go to donotreply.com, and realize that some of it is actually dangerous. The second way is to think, "This looks like a harmless failure, but we should avoid it anyway. No good can come of this." The first way protects you if you're clever; the second way always protects you.

"The first way protects you if you're clever; the second way always protects you." That's very much the other half of the security mindset. It's what this essay's author was doing by talking about AGI alignment that runs on whitelisting rather than blacklisting: you shouldn't assume you'll be clever about how fast the AGI system could discover capabilities, you should have a system that doesn't use not-yet-whitelisted capabilities even if they are discovered very suddenly.

If your AGI would hurt you if it gained total cosmic powers in one millisecond, that means you built a cognitive process that is in some sense trying to hurt you and failing only due to what you think is a lack of capability. This is *very bad* and you should be designing some other AGI system instead. AGI systems should never be running a search that will hurt you if the search comes up non-empty. You should not be trying to fix that by making sure the search comes up empty thanks to your clever shallow defenses closing off all the AGI's clever avenues for hurting you. You should fix that by making sure no search like that ever runs. It's a silly thing to do with computing power, and you should do something else with computing power instead.

Going back to ordinary computer security, if you try building a lock with seven keys hidden in different places, you are in some dimension pitting your cleverness against an adversary trying to read the keys. The person with security mindset doesn't want to rely on having to win the cleverness contest. An ordinary paranoid, somebody who can master the kind of default paranoia that lots of intelligent programmers have, will look at the Reply-To field saying donotreply@donotreply.com and think about the possibility of an adversary registering the donotreply.com domain. Somebody with security mindset thinks in assumptions rather than adversaries. "Well, I'm assuming that this reply email goes nowhere," they'll think, "but maybe I should design the system so that I don't need to fret about whether that assumption is true."

AMBER: Because as the truly great paranoid knows, what seems like a ridiculously improbable way for the adversary to attack sometimes turns out to not be so ridiculous after all.

CORAL: Again, that's a not-exactly-right way of putting it. When I don't set up an email to originate from donotreply@donotreply.com, it's not just because I've appreciated that an adversary registering donotreply.com is more probable than the novice imagines. For all I know, when a bounce email is sent to nowhere, there's all kinds of things that might happen! Maybe the way a bounced email works is that the email gets routed around to weird places looking for that address. I don't know, and I don't want to have to study it. Instead I'll ask: Can I make it so that a bounced email doesn't generate a reply? Can I make it so that a bounced email doesn't contain the text of the original message? Maybe I can query the email server to make sure it still has a user by that name before I try sending the message?—though there may still be

“vacation” autoresponses that mean I'd better control the replied-to address myself. If it would be very bad for somebody unauthorized to read this, maybe I shouldn't be sending it in plaintext by email.

AMBER: So the person with true security mindset understands that where there's one problem, demonstrated by what seems like a very unlikely thought experiment, there's likely to be more realistic problems that an adversary can in fact exploit. What I think of as weird improbable failure scenarios are canaries in the coal mine, that would warn a truly paranoid person of bigger problems on the way.

CORAL: Again that's not exactly right. The person with ordinary paranoia hears about donotreply@donotreply.com and may think something like, “Oh, well, it's not very likely that an attacker will actually try to register that domain, I have more urgent issues to worry about,” because in that mode of thinking, they're running around putting out things that might be fires, and they have to prioritize the things that are most likely to be fires.

If you demonstrate a weird edge-case thought experiment to somebody with security mindset, they don't see something that's more likely to be a fire. They think, “Oh no, my belief that those bounce emails go nowhere was FALSE!” The OpenBSD project to build a secure operating system has also, in passing, built an extremely robust operating system, because from their perspective any bug that potentially crashes the system is considered a critical security hole. An ordinary paranoid sees an input that crashes the system and thinks, “A crash isn't as bad as somebody stealing my data. Until you demonstrate to me that this bug can be used by the adversary to steal data, it's not *extremely* critical.” Somebody with security mindset thinks, “Nothing inside this subsystem is supposed to behave in a way that crashes the OS. Some section of code is behaving in a way that does not work like my model of that code. Who knows what it might do? The system isn't supposed to crash, so by making it crash, you have demonstrated that my beliefs about how this system works are false.”

AMBER: I'll be honest: It *has* sometimes struck me that people who call themselves security professionals seem overly concerned with what, to me, seem like very improbable scenarios. Like somebody forgetting to check the end of a buffer and an adversary throwing in a huge string of characters that overwrite the end of the stack with a return address that jumps to a section of code somewhere else in the system that does something the adversary wants. How likely is that *really* to be a problem? I suspect that in the real world, what's more likely is somebody making their password “password”. Shouldn't you be mainly guarding against that instead?

CORAL: You have to do both. This game is short on consolation prizes. If you want your system to resist attack by major governments, you need it to actually be pretty darned secure, gosh darn it. The fact that some users may try to make their password be “password” does not change the fact that you also have to protect against buffer overflows.

AMBER: But even when somebody with security mindset designs an operating system, it often still ends up with successful attacks against it, right? So if this deeper paranoia doesn't eliminate all chance of bugs, is it really worth the extra effort?

CORAL: If you don't have somebody who thinks this way in charge of building your operating system, it has *no* chance of not failing immediately. People with security mindset sometimes fail to build secure systems. People without security mindset

always fail at security if the system is at all complex. What this way of thinking buys you is a *chance* that your system takes longer than 24 hours to break.

AMBER: That sounds a little extreme.

CORAL: History shows that reality has not cared what you consider “extreme” in this regard, and that is why your Wi-Fi-enabled lightbulb is part of a Russian botnet.

AMBER: Look, I understand that you want to get all the fiddly tiny bits of the system exactly right. I like tidy neat things too. But let's be reasonable; we can't always get everything we want in life.

CORAL: You think you're negotiating with me, but you're really negotiating with Murphy's Law. I'm afraid that Mr. Murphy has historically been quite unreasonable in his demands, and rather unforgiving of those who refuse to meet them. I'm not advocating a policy to you, just telling you what happens if you don't follow that policy. Maybe you think it's not particularly bad if your lightbulb is doing denial-of-service attacks on a mattress store in Estonia. But if you do want a system to be secure, you need to do certain things, and that part is more of a law of nature than a negotiable demand.

AMBER: Non-negotiable, eh? I bet you'd change your tune if somebody offered you twenty thousand dollars. But anyway, one thing I'm surprised you're not mentioning more is the part where people with security mindset always submit their idea to peer scrutiny and then accept what other people vote about it. I do like the sound of that; it sounds very communitarian and modest.

CORAL: I'd say that's part of the ordinary paranoia that lots of programmers have. The point of submitting ideas to others' scrutiny isn't that hard to understand, though certainly there are plenty of people who don't even do that. If I had any original remarks to contribute to that well-worn topic in computer security, I'd remark that it's framed as advice to wise paranoids, but of course the people who need it even more are the happy innocents.

AMBER: Happy innocents?

CORAL: People who lack even ordinary paranoia. Happy innocents tend to envision ways that their system works, but not ask *at all* how their system might fail, until somebody prompts them into that, and even then they can't do it. Or at least that's been my experience, and that of many others in the profession.

There's a certain incredibly terrible cryptographic system, the equivalent of the Fool's Mate in chess, which is sometimes converged on by the most total sort of amateur, namely Fast XOR. That's picking a password, repeating the password, and XORing the data with the repeated password string. The person who invents this system may not be able to take the perspective of an adversary at all. *He* wants his marvelous cipher to be unbreakable, and he is not able to truly enter the frame of mind of somebody who wants his cipher to be breakable. If you ask him, “Please, *try* to imagine what could possibly go wrong,” he may say, “Well, if the password is lost, the data will be forever unrecoverable because my encryption algorithm is too strong; I guess that's something that could go wrong.” Or, “Maybe somebody sabotages my code,” or, “If you really insist that I invent far-fetched scenarios, maybe the computer spontaneously decides to disobey my programming.” Of course any competent ordinary paranoid asks the most skilled people they can find to look at a bright idea and try to shoot it down, because other minds may come in at a different angle or

know other standard techniques. But the other reason why we say “Don't roll your own crypto!” and “Have a security expert look at your bright idea!” is in hopes of reaching the many people who can't *at all* invert the polarity of their goals—they don't think that way spontaneously, and if you try to force them to do it, their thoughts go in unproductive directions.

AMBER: Like... the same way many people on the Right/Left seem utterly incapable of stepping outside their own treasured perspectives to pass the Ideological Turing Test of the Left/Right.

CORAL: I don't know if it's exactly the same mental gear or capability, but there's a definite similarity. Somebody who lacks ordinary paranoia can't take on the viewpoint of somebody who wants Fast XOR to be breakable, and pass that adversary's Ideological Turing Test for attempts to break Fast XOR.

AMBER: Can't, or won't? You seem to be talking like these are innate, untrainable abilities.

CORAL: Well, at the least, there will be different levels of talent, as usual in a profession. And also as usual, talent vastly benefits from training and practice. But yes, it has sometimes seemed to me that there is a kind of qualitative step or gear here, where some people can shift perspective to imagine an adversary that truly wants to break their code... or a reality that isn't cheering for their plan to work, or aliens who evolved different emotions, or an AI that doesn't *want* to conclude its reasoning with “And therefore the humans should live happily ever after”, or a fictional character who believes in Sith ideology and yet doesn't believe they're the bad guy.

It does sometimes seem to me like some people simply can't shift perspective in that way. Maybe it's not that they truly lack the wiring, but that there's an instinctive political off-switch for the ability. Maybe they're scared to let go of their mental anchors. But from the outside it looks like the same result: some people do it, some people don't. Some people spontaneously invert the polarity of their internal goals and spontaneously ask how their cipher might be broken and come up with productive angles of attack. Other people wait until prompted to look for flaws in their cipher, or they demand that you argue with them and wait for you to come up with an argument that satisfies them. If you ask them to predict themselves what you might suggest as a flaw, they say weird things that don't begin to pass your Ideological Turing Test.

AMBER: You do seem to like your qualitative distinctions. Are there better or worse ordinary paranoids? Like, is there a spectrum in the space between “happy innocent” and “true deep security mindset”?

CORAL: One obvious quantitative talent level within ordinary paranoia would be in how far you can twist your perspective to look sideways at things—the creativity and workability of the attacks you invent. Like these examples Bruce Schneier gave:

> Uncle Milton Industries has been selling ant farms to children since 1956. Some years ago, I remember opening one up with a friend. There were no actual ants included in the box. Instead, there was a card that you filled in with your address, and the company would mail you some ants. My friend expressed surprise that you could get ants sent to you in the mail.
>
> I replied: “What's really interesting is that these people will send a tube of live ants to anyone you tell them to.”

> Security requires a particular mindset. Security professionals—at least the good ones—see the world differently. They can't walk into a store without noticing how they might shoplift. They can't use a computer without wondering about the security vulnerabilities. They can't vote without trying to figure out how to vote twice. They just can't help it.
>
> SmartWater is a liquid with a unique identifier linked to a particular owner. “The idea is for me to paint this stuff on my valuables as proof of ownership,” I wrote when I first learned about the idea. “I think a better idea would be for me to paint it on your valuables, and then call the police.”
>
> Really, we can't help it.
>
> This kind of thinking is not natural for most people. It's not natural for engineers. Good engineering involves thinking about how things can be made to work; the security mindset involves thinking about how things can be made to fail...
>
> I've often speculated about how much of this is innate, and how much is teachable. In general, I think it's a particular way of looking at the world, and that it's far easier to teach someone domain expertise—cryptography or software security or safecracking or document forgery—than it is to teach someone a security mindset.

To be clear, the distinction between “just ordinary paranoia” and “all of security mindset” is my own; I think it's worth dividing the spectrum above the happy innocents into two levels rather than one, and say, “This business of looking at the world from weird angles is only half of what you need to learn, and it's the easier half.”

AMBER: Maybe Bruce Schneier himself doesn't grasp what you mean when you say “security mindset”, and you've simply stolen his term to refer to a whole new idea of your own!

CORAL: No, the thing with not wanting to have to reason about whether somebody might someday register “donotreply.com” and just fixing it regardless—a methodology that doesn't trust you to be clever about which problems will blow up—that's definitely part of what existing security professionals mean by “security mindset”, and it's definitely part of the second and deeper half. The only unconventional thing in my presentation is that I'm factoring out an intermediate skill of “ordinary paranoia”, where you try to parry an imagined attack by encrypting your password file and hiding the encryption key in a separate section of filesystem code. Coming up with the idea of hashing the password file is, I suspect, a qualitatively distinct skill, invoking a world whose dimensions are your own reasoning processes and not just object-level systems and attackers. Though it's not polite to say, and the usual suspects will interpret it as a status grab, my experience with other reflectivity-laden skills suggests this may mean that many people, possibly including you, will prove unable to think in this way.

AMBER: I indeed find that terribly impolite.

CORAL: It may indeed be impolite; I don't deny that. Whether it's untrue is a different question. The reason I say it is because, as much as I want ordinary paranoids to *try* to reach up to a deeper level of paranoia, I want them to be aware that it might not prove to be their thing, in which case they should get help and then listen to that help. They shouldn't assume that because they can notice the chance to have ants mailed to people, they can also pick up on the awfulness of donotreply@donotreply.com.

AMBER: Maybe you could call that “deep security” to distinguish it from what Bruce Schneier and other security professionals call “security mindset”.

CORAL: “Security mindset” equals “ordinary paranoia” plus “deep security”? I'm not sure that's very good terminology, but I won't mind if you use the term that way.

AMBER: Suppose I take that at face value. Earlier, you described what might go wrong when a happy innocent tries and fails to be an ordinary paranoid. What happens when an ordinary paranoid tries to do something that requires the deep security skill?

CORAL: They believe they have wisely identified bad passwords as the real fire in need of putting out, and spend all their time writing more and more clever checks for bad passwords. They are very impressed with how much effort they have put into detecting bad passwords, and how much concern they have shown for system security. They fall prey to the standard cognitive bias whose name I can't remember, where people want to solve a problem using one big effort or a couple of big efforts and then be done and not try anymore, and that's why people don't put up hurricane shutters once they're finished buying bottled water. Pay them to “try harder”, and they'll hide seven encryption keys to the password file in seven different places, or build towers higher and higher in places where a successful adversary is obviously just walking around the tower if they've gotten through at all. What these ideas have in common is that they are in a certain sense “shallow”. They are mentally straightforward as attempted parries against a particular kind of envisioned attack. They give you a satisfying sense of fighting hard against the imagined problem—and then they fail.

AMBER: Are you saying it's *not* a good idea to check that the user's password isn't “password”?

CORAL: No, shallow defenses are often good ideas too! But even there, somebody with the higher skill will try to look at things in a more systematic way; they know that there are often deeper ways of looking at the problem to be found, and they'll try to find those deep views. For example, it's extremely important that your password checker does *not* rule out the password “correct horse battery staple” by demanding the password contain at least one uppercase letter, lowercase letter, number, and punctuation mark. What you really want to do is measure password entropy. Not envision a failure mode of somebody guessing “rainbow”, which you will cleverly balk by forcing the user to make their password be “rA1nbow!” instead.

You want the password entry field to have a checkbox that allows showing the typed password in plaintext, because your attempt to parry the imagined failure mode of some evildoer reading over the user's shoulder may get in the way of the user entering a long or high-entropy password. And the user is perfectly capable of typing their password into that convenient text field in the address bar above the web page, so they can copy and paste it—thereby sending your password to whoever tries to do smart lookups on the address bar. If you're really that worried about some evildoer reading over somebody's shoulder, maybe you should be sending a confirmation text to their phone, rather than forcing the user to enter their password into a nearby text field that they can actually read. Obscuring one text field, with no off-switch for the obscuration, to guard against this one bad thing that you imagined happening, while managing to step on your own feet in other ways and not even really guard against the bad thing; that's the peril of shallow defenses.

An archetypal character for “ordinary paranoid who thinks he's trying really hard but is actually just piling on a lot of shallow precautions” is Mad-Eye Moody from the *Harry Potter* series, who has a whole room full of Dark Detectors, and who also ends up locked in the bottom of somebody's trunk. It seems Mad-Eye Moody was too busy buying one more Dark Detector for his existing room full of Dark Detectors, and he didn't invent precautions deep enough and general enough to cover the unforeseen attack vector “somebody tries to replace me using Polyjuice”.

And the solution isn't to add on a special anti-Polyjuice potion. I mean, if you happen to have one, great, but that's not where most of your trust in the system should be coming from. The first lines of defense should have a sense about them of depth, of generality. Hashing password files, rather than hiding keys; thinking of how to measure password entropy, rather than requiring at least one uppercase character.

AMBER: Again this seems to me more like a quantitative difference in the cleverness of clever ideas, rather than two different modes of thinking.

CORAL: Real-world categories are often fuzzy, but to me these seem like the product of two different kinds of thinking. My guess is that the person who popularized demanding a mixture of letters, cases, and numbers was reasoning in a different way than the person who thought of measuring password entropy. But whether you call the distinction qualitative or quantitative, the distinction remains. Deep and general ideas—the kind that actually simplify and strengthen the edifice of reasoning supporting the system's safety—are invented more rarely and by rarer people. To build a system that can resist or even slow down an attack by multiple adversaries, some of whom may be smarter or more experienced than ourselves, requires a level of professionally specialized thinking that isn't reasonable to expect from every programmer—not even those who can shift their minds to take on the perspective of a single equally-smart adversary. What you should ask from an ordinary paranoid is that they appreciate that deeper ideas exist, and that they try to learn the standard deeper ideas that are already known; that they know their own skill is not the upper limit of what's possible, and that they ask a professional to come in and check their reasoning. And then actually listen.

AMBER: But if it's possible for people to think they have higher skills and be mistaken, how do you know that *you* are one of these rare people who *truly* has a deep security mindset? Might your high opinion of yourself just be due to the Dunning-Kruger effect?

CORAL: ... Okay, that reminds me to give another caution.

Yes, there will be some innocents who can't believe that there's a talent called “paranoia” that they lack, who'll come up with weird imitations of paranoia if you ask them to be more worried about flaws in their brilliant encryption ideas. There will also be some people reading this with severe cases of social anxiety and underconfidence. Readers who *are* capable of ordinary paranoia and even security mindset, who might not try to develop these talents, because they are terribly worried that they might just be one of the people who only imagine themselves to have talent. Well, if you think you can feel the distinction between deep security ideas and shallow ones, you should at least try now and then to generate your own thoughts that resonate in you the same way.

AMBER: But won't that attitude encourage overconfident people to think they can be paranoid when they actually can't be, with the result that they end up too impressed with their own reasoning and ideas?

CORAL: I strongly suspect that they'll do that regardless. You're not actually promoting some kind of collective good practice that benefits everyone, just by personally agreeing to be modest. The overconfident don't care what you decide. And if you're not just as worried about underestimating yourself as overestimating yourself, if your fears about exceeding your proper place are asymmetric with your fears about lost potential and foregone opportunities, then you're probably dealing with an emotional issue rather than a strict concern with good epistemology.

AMBER: If somebody does have the talent for deep security, then, how can they train it?

CORAL: … That's a hell of a good question. Some interesting training methods have been developed for ordinary paranoia, like classes whose students have to figure out how to attack everyday systems outside of a computer-science context. One professor gave a test in which one of the questions was "What are the first 100 digits of pi?"—the point being that you need to find some way to cheat in order to pass the test. You should train that kind of ordinary paranoia first, if you haven't done that already.

AMBER: And then what? How do you graduate to deep security from ordinary paranoia?

CORAL: … Try to find more general defenses instead of parrying particular attacks? Appreciate the extent to which you're building ever-taller versions of towers that an adversary might just walk around? Ugh, no, that's too much like ordinary paranoia—especially if you're starting out with just ordinary paranoia. Let me think about this.

…

Okay, I have a screwy piece of advice that's probably not going to work. Write down the safety-story on which your belief in a system's security rests. Then ask yourself whether you actually included all the empirical assumptions. Then ask yourself whether you actually believe those empirical assumptions.

AMBER: So, like, if I'm building an operating system, I write down, "Safety assumption: The login system works to keep out attackers"—

CORAL: *No!*

Uh, no, sorry. As usual, it seems that what I think is "advice" has left out all the important parts anyone would need to actually do it.

That's not what I was trying to handwave at by saying "empirical assumption". You don't want to assume that parts of the system "succeed" or "fail"—that's not language that should appear in what you write down. You want the elements of the story to be strictly factual, not... value-laden, goal-laden? There shouldn't be reasoning that explicitly mentions what you want to have happen or not happen, just language neutrally describing the background facts of the universe. For brainstorming purposes you might write down "Nobody can guess the password of any user with dangerous privileges", but that's just a proto-statement which needs to be refined into more basic statements.

AMBER: I don't think I understood.

CORAL: "Nobody can guess the password" says, "I believe the adversary will fail to guess the password." Why do you believe that?

AMBER: I see, so you want me to refine complex assumptions into systems of simpler assumptions. But if you keep asking “why do you believe that” you'll eventually end up back at the Big Bang and the laws of physics. How do I know when to stop?

CORAL: What you're trying to do is reduce the story past the point where you talk about a goal-laden event, “the adversary fails”, and instead talk about neutral facts underlying that event. For now, just answer me: Why do you believe the adversary fails to guess the password?

AMBER: Because the password is too hard to guess.

CORAL: The phrase “too hard” is goal-laden language; it's your own desires for the system that determine what is “too hard”. Without using concepts or language that refer to what you want, what is a neutral, factual description of what makes a password too hard to guess?

AMBER: The password has high-enough entropy that the attacker can't try enough attempts to guess it.

CORAL: We're making progress, but again, the term “enough” is goal-laden language. It's your own wants and desires that determine what is “enough”. Can you say something else instead of “enough”?

AMBER: The password has sufficient entropy that—

CORAL: I don't mean find a synonym for “enough”. I mean, use different concepts that aren't goal-laden. This will involve changing the meaning of what you write down.

AMBER: I'm sorry, I guess I'm not good enough at this.

CORAL: Not yet, anyway. Maybe not ever, but that isn't known, and you shouldn't assume it based on one failure.

Anyway, what I was hoping for was a pair of statements like, “I believe every password has at least 50 bits of entropy” and “I believe no attacker can make more than a trillion tries total at guessing any password”. Where the point of writing “I believe” is to make yourself pause and question whether you actually believe it.

AMBER: Isn't saying no attacker “can” make a trillion tries itself goal-laden language?

CORAL: Indeed, that assumption might need to be refined further via why-do-I-believe-that into, “I believe the system rejects password attempts closer than 1 second together, I believe the attacker keeps this up for less than a month, and I believe the attacker launches fewer than 300,000 simultaneous connections.” Where again, the point is that you then look at what you've written and say, “Do I really believe that?” To be clear, sometimes the answer will be “Yes, I sure do believe that!” This isn't a social modesty exercise where you show off your ability to have agonizing doubts and then you go ahead and do the same thing anyway. The point is to find out what you believe, or what you'd need to believe, and check that it's believable.

AMBER: And this trains a deep security mindset?

CORAL: … Maaaybe? I'm wildly guessing it might? It may get you to think in terms of stories and reasoning and assumptions alongside passwords and adversaries, and that puts your mind into a space that I think is at least part of the skill.

In point of fact, the real reason the author is listing out this methodology is that he's currently trying to do something similar on the problem of aligning Artificial General Intelligence, and he would like to move past “I believe my AGI won't want to kill anyone” and into a headspace more like writing down statements such as “Although the space of potential weightings for this recurrent neural net does contain weight combinations that would figure out how to kill the programmers, I believe that gradient descent on loss function *L* will only access a result inside subspace *Q* with properties *P*, and I believe a space with properties *P* does not include any weight combinations that figure out how to kill the programmer.”

Though this itself is not really a reduced statement and still has too much goal-laden language in it. A realistic example would take us right out of the main essay here. But the author does hope that practicing this way of thinking can help lead people into building more solid stories about robust systems, if they already have good ordinary paranoia and some fairly mysterious innate talents.

To be continued in: **Security Mindset and the Logistic Success Curve**

Cooperative model knob-turning

(Cross-posted from my medium channel)

When you are trying to understand something by yourself, a useful skill to check your grasp on the subject is to try out the moving parts of your mental model and see if you can simulate the resulting changes.

Suppose you want to learn how a rocket works. At the bare minimum, you should be able to calculate the speed of the rocket given the time past launch. But can you tell what happens if Earth gravity was stronger? Weaker? What if the atmosphere had no oxygen? What if we replaced the fuel with Diet Coke and Mentos?

To really understand something, it's not enough to be able to predict the future in a normal, expected, *ceteris paribus* scenario. You should also be able to predict what happens when several variables are changed is several ways, or, at least, point to which calculations need to be run to arrive at such a prediction.

Douglas Hofstadter and Daniel Dennett call that "turning the knobs". Imagine your model as a box with several knobs, where each knob controls one aspect of the modeled system. You don't have to be able to turn all the possible knobs to all possible values and still get a sensible, testable and correct answer, but the more, the better.

Doug and Dan apply this approach to thought experiments and intuition pumps, as a way to explore possible answers to philosophical questions. In my experience, this skill is also effective when applied to real world problems, notably when trying to understand something that is being explained by someone else.

In this case, you can run this knob-turning check cooperatively with the other person, which makes it way more powerful. If someone says “X+Y = Z” and “X+W = Z+A”, it’s not enough to mentally turn the knobs and calculate “X+Y+W = Z+A+B”. You should do that, then actually ask the explainer “Hey, let me see if I get what you mean: for example, X+Y+W would be Z+A+B”?

This cooperative model knob-turning has been useful to me in many walks of life, but the most common and mundane application is helping out people at work. In that context, I identify six effects which make it helpful:

1) Communication check: maybe you misunderstood and actually X+W = Z-A

This is useful overall, but very important if someone uses metaphor. Some metaphors are clearly vague and people will know that and avoid them in technical explanations. But some metaphors seem really crisp for some people but hazy to others, or worse, very crisp to both people, but with different meanings! So take every metaphor as an invitation to interactive knob-turning.

To focus on communication check, try rephrasing their statements, using different words or, if necessary, very different metaphors. You can also apply a theory in different contexts, to see if the metaphors still apply.

For example, if a person talks about a computer system as if it were a person, I might try to explain the same thing in terms of a group of trained animals, or a board of directors, or dominoes falling.

2) Self-check: correct your own reasoning (maybe you understood the correct premises, but made a logical mistake during knob turning)

This is useful because humans are fallible, and two (competent) heads are less likely to miss a step in the reasoning dance than one.

Also, when someone comes up and asks something, you'll probably be doing a context-switch, and will be more likely to get confused along the way. The person asking usually has more local context than you in the specific problem they are trying to solve, even if you have more context on the surrounding matters, so they might be able to spot your error more quickly than yourself.

Focus on self-check means double checking any intuitive leaps or tricky reasoning you used. Parts of your model that do not have a clear step-by-step explanation have priority, and should be tested against another brain. Try to phrase the question in a way that makes your intuitive answer look *less* obvious.

For example, suppose you are helping someone build a group chat system where the order of messages matter, i.e. if Alice asks Bob something and he answers, you don't want Charlie to see Bob's answer before he sees Alice's question. But intuitively it looks like that doesn't matter for messages in different chat rooms. So you ask: *"I'm not sure about this, but I think we can relax the requirement for message order if the messages are in different chatrooms, right?"*

3) Other-check: help the other person to correct inferential errors they might have made

The converse of self-checking. Sometimes fresh eyes with some global context can see reasoning errors that are hidden to people who are very focused on a task for too long.

To focus on other-check, ask about conclusions that follow from your model of the situation, but seem unintuitive to you, or required tricky reasoning. It's possible that your friend also found them unintuitive, and that might have lead them to a jump to the opposite direction.

For example, in computers, and moreso in distributed systems, clocks don't always work the we naively expect. So if I think that's relevant to the problem at hand, I could ask: *"For this system to work correctly, it seems that the clocks have to be closely synchronized, right? If the clocks are off by much, we could have a difference around midnight. Did you take that into account?"*

Perhaps you successfully understand what was said, and the model you built in your head fits the communicated data. But that doesn't mean it is the same model that the other person has in mind! In that case, your knob-turning will get you a result that's inconsistent with what they expect.

4) Alternative hypothesis generation: If they cannot refute your conclusions, you have shown them a possible model they had not yet considered, in which case it will also point in the direction of more research to be made

This is incredibly relevant when someone asks for help debugging. If they can't find the root cause of a bug, it must be because they are missing something. Either they have derived a mistaken conclusion from the data, or they are missing relevant data, or they've made an inferential error from those conclusions. The first two cases are where proposing a new model helps, highlighting ambiguities and pointing to new focuses of empirical investigation (the third case is solved by other-checking).

Maybe they read the system logs, saw that a task was initiated (for example, writing a file), and assumed it was completed successfully, but perhaps it wasn't. In that case, you can tell them to check for a record indicating the completion of the task, or the absence of such a record.

To boost this effect, look for data that you strongly expect to exist and confirm your model, but wasn't shown by your colleague. Either the data doesn't exist and your model is wrong, or they actually didn't look into it, and they should. You can prioritize areas where you think they have relatively less context, skill or experience.

For example: *"Ok, so if the database went down, we should've seen all requests failing in that time range; but if it was a network instability, we should have random requests failing and others succeeding. Which one was it?"*

5) Filling gaps in context: If they show you data that contradicts your model, well, you get more data and improve your understanding

This is very important when you have much less context than the other person. The larger the difference in context, the more likely that there's some important piece of information that you don't have, but that they take for granted.

The point here isn't that there is something you don't know. There are lots and lots of things you and your colleague don't know. If there's something they know that you don't, they'll probably fill you in when asking the question.

The point is that they will tell you something *only if they realize you don't know it yet*. But people will expect short inferential distances, underestimate the difference in

context, and forget to tell you stuff because it just seems obvious to them that you know all that.

A good way to fill gaps is to ask about the parts of your model which you are more uncertain about, to find out if they can help you build a clearer image. You can also extrapolate and make a wild guess, which you don't really expect to be right.

For example: *"How does the network works on this datacenter? Do we have a single switch so that, if it fails, all connections go down? Or are those network interfaces all virtualized anyway?"*

6) Finding new ideas: If everybody understands one another, and the models are correct, knob-turning will lead to new conclusions (if they hadn't turned those specific knobs on the problem yet)

This is the whole point of having the conversation, to help someone figure something out they haven't already. But even if the specific new conclusion you arrive when knob-turning isn't directly relevant to the current question, it may end up shining light on some part of the other person's model that they couldn't see yet.

This effect is general and will happen gradually as both your and the other person's models improve and converge. The goal is to get all obstacles out of the way so you can just move forward and find new ideas and solutions.

The more context and relevant skill your colleague has, the lower the chance that they missed some crucial piece of data and have a mistaken model (or, if they do, you probably won't be able to figure that out without putting in serious effort). So when talking to more skilled or experienced people, you can focus more in replicating the model from their mind to yours (**communication check** and **self-check**).

Conversely, when talking to less skilled people, you should focus more on errors they might have made, or models they might not have considered, or data they may need to collect (**other-check** and **alternative hypothesis generation**).

Filling gaps depends more on differences of communication style and local context, so I don't have a person-based heuristic.

Preferences over non-rewards

In this penultimate post on "learning human values" series, I just want to address some human values/preferences/rewards that don't fit neatly into the (p, R) model where p in the planning algorithm and R the actual reward.

Preferences over preferences and knowledge

Most people have preferences over their own preferences - and that of others. For example, consider someone who has an incorrect religious faith. They might believe something like:

"I want to always continue believing. I flinch away from certain sceptical arguments, but I'm sure my deity would protect me from doubt if I ever decided to look into them".

Hope this doesn't sound completely implausible for someone. Here they have beliefs, preferences over their future beliefs, and beliefs over their future beliefs. This doesn't seem to be able to be easily captured in the (p, R) framework. We can also see that asking them equivalent questions "Do you want to doubt your deity?" and "Do you want to learn the truth?" will get very different answers.

But it's not just theism, an example which is too easy to pick on. I have preferences over knowledge, for instance, as do most people. I would prefer that people had accurate information, for instance. I would also prefer that, when choosing between possible formalisations of preferences, people went with the less destructive and less self-destructive options. These are not overwhelmingly strong preferences, but they certainly exist.

Aliefs

Consider the following scenario: someone believes that roller-coasters are perfectly safe, but enjoys riding them for the feeling of danger they give them. It's clear that the challenge here is not reconciling the belief of safety with the alief of danger (which is simple: roller-coasters are safe), but to somehow transform the feeling of danger into another form that keeps the initial enjoyment.

Tribalism and signalling

The theism argument might suggest that tribalism will be a major problem, as various groups pressure adherents to conform to certain beliefs and preferences.

But actually that need not be such a problem. It's clear that there is a strong desire to remain part of that group (or, sometimes, just of a group). Once that desire is identified, all the rest become instrumental - the human will either do the actions that are needed to remain part of the group, without needing to change their beliefs or preference (just because evolution doesn't allow us to separate those two easily, doesn't mean an AI can't help us do it), or will rationally sacrifice beliefs and preferences to the cause of remaining part of the group.

Most signalling cases can be dealt with in the same way. So, though tribalism is a major reason people can end up with contingent preferences, it doesn't in itself pose problems to the (p, R) model.

Personal identity

The problem of personal identity is a tricky one. I would like to remain alive, happy, curious, having interesting experience, doing worthwhile and varied activities, etc...

Now, this is partially preferences about future preferences, but there's the implicit identity: I want this to happen to *me*. Even when I'm being altruistic, I want these experiences to happen to *someone*, not just to happen in some abstract sense.

But the concept of personal identity is a complicated one, and it's not clear if it can be collapsed easily into the (p, R) format.

"You're not the boss of me!"

Finally, even if personal identity is defined, it remains the case that people can judge different situations depending on how that situation is achieved. Being forced or manipulated into a situation will make them resent it much more than if they reach it through "natural" means. Of course, what counts as acceptable and unacceptable manipulations change, is filled with biases, inconsistencies, and incorrect beliefs (in my experience, far too many people think themselves immune to advertising, for instance).

Caring about derivatives rather than positions

People react strongly to situations getting worse of better, not so much to the absolute quality of the situation.

Values that don't make sense out of context

AIs would radically reshape the world and society. And yet humans have deeply held values that only make sense in narrow contexts - sometimes, they already no longer make sense. For instance, in my opinion, of the five categories in moral foundations theory, one no longer makes sense and three only make partial sense (and it seems to me that having these values in a world where it's literally impossible to satisfy them, is part of the problem people have with the modern world):

- Care: cherishing and protecting others. This seems to me the strongest foundation; care remains well defined today, most especially in the negative "protect people from harm" sense.
- Purity: abhorrence for disgusting things, foods, actions. This seems the weakest foundation. Our ancestral instincts of disgust for food and people are no longer correlated with actual danger, or with anything much. Disgust is the easiest value to argue against, and the hardest to defend, because it provokes such

strong feelings but the boundaries drawn around the objects of disgust make no sense.
- Fairness: rendering justice according to shared rules. Fairness and equality make only partial sense in today's world. It seems impossible to ensure that every interaction is fair, and that everyone gets their just desert (whatever that means) or gets the same opportunities. But two subcategories do exists: legal rights fairness/equality, and financial fairness/equality. Modern societies achieve the first to some extent, and make attempts at the second.
- Authority: submitting to tradition and legitimate authority. This also makes partial sense. Traditions is a poor guide in many situations, and the source of authority doesn't simplify real problems or guarantee solutions (which is the main reasons that dictators are not generally any better at solving problems). As with fairness, the subcategory of legal authority is used extensively in the world today.
- Loyalty: standing with your group, family, nation. This value is weak, and may end up further weakening, down to the level of purity. There are basically too many positive sum interactions in today's world. The benefits of trade and interacting with those outside your ingroup, are huge. Legally, most of loyalty is actually forbidden - we don't have laws encouraging nepotism, rather the opposite.

This can be seen as a subset of the whole "underdefined human values", but it could also be seen as an argument for preserving or recreating certain contexts, in which these values make sense.

A more complex format needed

These are just some of the challenges to the (p, R) format, and there are certainly others. It's not clear how much that format needs to complicated in order to usefully model all these extra types of preferences.

Competitive Truth-Seeking

In many domains, you can get better primarily by being correct more frequently. If you're managing a team, or trying to improve your personal relationships, it's very effective to improve your median decision. The more often you're right, the better you'll do, so people often implicitly optimize their success rate.

But what about competitive domains, like prediction markets, investing, or hiring?

One of the most common mistakes I see people make is trying to apply non-competitive intuitions to competitive dynamics. But the winning strategies are *very* different! It is not enough to be right – others must be wrong. Rather than having a high success rate across all situations, you want to find an edge in some situations, and bet heavily when you find a mistake in the consensus.

For example, say you're a startup trying to hire software engineers, and your process finds two excellent candidates – Alex and Bob. You make both of them offers at market rate for excellent software engineers. Alex knows how to interview well, so every company believes (correctly) that he's excellent, while Bob is a bad interviewer, and most of your competitors think he's merely good. As a result, there might be a 10% chance Alex accepts your offer in this situation, and a 50% chance Bob does.

Assuming a similar rate of Alexes and Bobs in the world, this means your process should be optimized almost entirely for finding Bobs – if you miss a few Alexes to find another Bob, that's totally worth it! But from the outside, this will look crazy – you're passing up obviously great candidates in order to specifically find people who aren't obviously great.

Furthermore, this means that copying mainstream interviewing strategies is one of the worst things you can do – if you only get points for beating consensus predictions, then matching them will get you a 0.

The next time you find yourself thinking about one of these situations, try to figure out whether it's competitive or non-competitive, and whether you should be optimizing for median predictions or edge.

'X is not about Y' is not about psychology

As Robin Hanson says:

> *Food isn't about Nutrition*
> *Clothes aren't about Comfort*
> *Bedrooms aren't about Sleep*
> *Marriage isn't about Romance*
> *Talk isn't about Info*
> *Laughter isn't about Jokes*
> *Charity isn't about Helping*
> *Church isn't about God*
> *Art isn't about Insight*
> *Medicine isn't about Health*
> *Consulting isn't about Advice*
> *School isn't about Learning*
> *Research isn't about Progress*
> *Politics isn't about Policy*

As I understand it, the prototypical form of Hansonian skepticism is to look at an X which is overtly about Y, and argue that the actual behavior of humans engaging in X is not consistent with caring about Y. Instead, the Hansonian cynic concludes, some unspoken motives must be at work; most often, status motives. He has a new book with Kevin Silmer which (by the sound of it) covers a lot of examples.

For him, the story of what's going on in cases like this has a lot to do with evolutionary psychology -- humans are adapted to play these games about what we want, even if we don't know it. Our brains keep the conscious justifications seperate from the hidden motives, so that we can explain ourselves in public while still going after private goals.

I'm not denying that this plays a role, but I think there's a more general phenomenon here which has nothing to do with humans or evolutionary psychology.

In economics, models based on an assumption of rational agents are fairly successful at making predictions, even though we know individual humans are far from rational. As long as the market contains inefficiencies which can be exploited, *someone* can jump in and exploit them. Not everyone needs to be rational, and no one person needs to be rational about everything, for the market to appear rational.

Similarly, the tails come apart whether or not X and Y have to do with humans. Someone might score high on calibration, but that doesn't make their strongest beliefs the most trustworthy -- indeed, they're the ones we ought to downgrade the most, even if we know nothing about human psychology and reason only from statistics. This alone is enough for us not to be too surprised at some of the "X is not about Y" examples -- we might expect explicit justification to be mostly good explanations for behavior, but by default, this shouldn't make us think that the correltion will be perfect. And if we optimize for justifiability, we expect the correlation to decrease.

Almost any problem has to be solved by coming up with proxies for progress. As Goodhart's Law notes, that these proxies will have errors which naturally come to be

exploited, simply by virtue of individuals in the system following incentive gradients. "Food is not about nutrition" because evolution found an imperfect proxy for nutritive value, and the modern economy optimized food to cheat the biological system. Organizations fall into lipservice to lost purposes because they set up incentive structures which are good enough initially, but eventually get exploited. You can't usually call out such failures in a way the system will recognize, because the exploitation will happen in a way that's crafted to the blind spots -- and there's no need for the exploiters to be aware, consciously or subconsciously, of what they're doing. Any system will tend to optimize for seeming over being, because seeming is what's directly available to optimize. It's the principal-agent problem all the way down, from national elections to a person's internal motivation system. Call it wireheading. Call it Goodhart. It's all the same cluster of dysfunction.

Machine learning researchers understand this phenomen in their domain -- they call it overfitting, and fight it with regularization. In machine learning, you want predictive accuracy. But, you can't directly optimize for accuracy on the data you have available -- you'll get something which is very bad at predicting future data. The solution to this is *regularization*: you optimize a combination of predictive accuracy and a simplicity measure. This creates some "drag" on the optimization, so that it doesn't over-optimize predictive accuracy on the available data at the expense of future accuracy. This works extremely well, considering that it doesn't necessarily introduce any new information to correlate things with reality better (though, in the Bayesian case, the regulariser is a prior probability which may contain real information about which things are more likely). If we had a general theory of anti-Goodharting which worked as well as this, it would seem to have broad implications for group coordination and societal problems.

This pessimism shouldn't be pushed too hard; usually, you're better off coming up with *some kind of* proxy measurement for your goals, so that you can shift your strategies in the face of measured success and failure. Your taste in food isn't *that* bad a proxy for what's good for you. Profit isn't *that* bad a proxy for providing value. Scientific consensus isn't *that* bad a proxy for truth. And so on. But keep in mind: when an optimization process is based on a proxy, there may be systematic failures. X is not about Y.

Open thread, November 13 - November 20, 2017

Given that weekly open threads are a feature of the old LessWrong, let's try them here as well.

If it's worth saying, but not worth its own post, then it goes here.

Clarify Your No's

As humans, we are often too quick to say that something is impossible and move on. Often it pays to double-check - what exactly have we shown to be impossible and does this really matter? Sometimes we'll find it is actually possible, sometimes it'll just give us another idea, but in both these cases it is useful.

Example 1: **Can a computer work without power**?

- Suppose you've been teleported into the past, but you would to be able to solve a calculation. You have no way of generating electricity. Just because an e*lectronic computer* like the one you buy at the store needs power, it doesn't mean that you can't build a mechanical computer that doesn't need power. (clarifying computer)
- Suppose you're making a computer. Surely, it can't be expected to do anything when the power is off? But what about if it is e-ink, then it can display something static. (clarifying work)
- Suppose you want to give children laptop with educational programs, but the town has no sources of electricity. You invent the OLPC laptops, which are powered using handcranks. (clarifying without power)

In each of these cases, it would have been very easy to think, "Computers need power" there is nothing that I can do, when this wasn't the case. Not that we ended up clarifying each phrase - "computer", "work", "without power"

Example 2: Barber Paradox (wiki)

The barber says that they shave every man in the town, but no man shaves themselves.

If you haven't heard this before, this brainteaser might be hard or it might be obvious. Regardless, we can make it even easer by clarifying each phrase.

- What does "in the town?" mean? Maybe they mean everyone who lives in the town and doesn't include people who work in the town.
- What if the barber is not a man? What if they are a women?

The second is the proper solution, but it comes out very easily once we start clarifying exactly what is impossible?

Example 3: **Being attractive**

Suppose a person wants to be attractive. Suppose they find out that being attractive requires them to be tall and have lots of muscles because they live in Shallow Land. Unfortunately, they are short and can't gain muscles no matter how much they train. Perhaps, they even become depressed because of this.

On the other hand, what is meant by "attractive"? What do they actually want? Perhaps they don't mean "attractive to most people", but "attractive to a reasonable subset of people"? So perhaps they will never have most people seeing them as attractive, but they could become attractive within the intelligentsia sub-culture. Or perhaps they only need to be attractive enough to find someone to marry? The initial no turned out to be less important than it seemed.

Take-aways

One of the key lessons here is that language is slippery. It's very easy think think, "That's impossible, let's move on", without realising that the thing that is impossible is actually rather limited.

Even when it is actually impossible, we can generate ideas by trying to discover what is the closest that we can get to achieving something that is impossible.

The main technique to accomplish this, it to attempt clarify each word or phrase. If you do this, then ideas drop out pretty quickly.

The Happy Dance Problem

[Cross-posted from IAFF.]

Since the invention of logical induction, people have been trying to figure out what logically updateless reasoning could be. This is motivated by the idea that, in the realm of Bayesian uncertainty (IE, empirical uncertainty), updateless decision theory is the simple solution to the problem of reflective consistency. Naturally, we'd like to import this success to logically uncertain decision theory.

At a research retreat during the summer, we realized that updateless decision theory wasn't so easy to define even in the seemingly simple Bayesian case. A possible solution was written up in Conditioning on Conditionals. However, that didn't end up being especially satisfying.

Here, I introduce the happy dance problem, which more clearly illustrates the difficulty in defining updateless reasoning in the Bayesian case. I also outline Scott's current thoughts about the correct way of reasoning about this problem.

(Ideas here are primarily due to Scott.)

The Happy Dance Problem

Suppose an agent has some chance of getting a pile of money. In the case that the agent gets the pile of money, it has a choice: it can either do a happy dance, or not. The agent would rather not do the happy dance, as it is embarrassing.

I'll write "you get a pile of money" as M, and "you do a happy dance" as H.

So, the agent has the following utility function:

- U(¬M) = $0
- U(M & ¬H) = $1000
- U(M & H) = $900

A priori, the agent assigns the following probabilities to events:

- P(¬M) = .5
- P(M & ¬H) = .1
- P(M & H) = .4

IE, the agent expects itself to do the happy dance.

Conditioning on Conditionals

In order to make an updateless decision, we need to condition on the policy of dancing, and on the policy of not dancing. How do we condition on a policy? We could change the problem statement by adding a policy variable and putting in the conditional probabilities of everything given the different policies, but this is just cheating: in order to fill in those conditional probabilities, you need to already know

how to condition on a policy. (This simple trick seems to be what kept us from noticing that UDT isn't so easy to define in the Bayesian setting for so long.)

A naive attempt would be to condition on the material conditional representing each policy, M⊃H and M⊃¬H. This gets the wrong answer. The material conditional simply rules out the one outcome inconsistent with the policy.

Conditioning on M⊃H, we get:

- P(¬M) = .555
- P(M & H) = .444

For an expected utility of $400.

Conditioning on M⊃¬H, we get:

- P(¬M) = .833
- P(M & ¬H) = .166

For an expected utility of $166.66.

So, to sum up, the agent thinks it should do the happy dance because refusing to do the happy dance makes worlds where it gets the money less probable. This doesn't seem right.

Conditioning on Conditionals solved this by sending the *probabilistic*conditional P(H|M) to one or zero to represent the effect of a policy, rather than using the material conditional. However, this approach is unsatisfactory for a different reason.

Happy dance is similar to Newcomb's problem with a transparent box (where Omega judges you on what you do when you see the full box): doing the dance is like one-boxing. Now, the correlation between doing the dance and getting the pile of money comes from Omega rather than just being part of an arbitrary prior. But, sending the conditional probability of one-boxing upon seeing the money to one doesn't make the world where the pile of money appears any more probable. So, this version of updateless reasoning gets transparent-box Newcomb wrong. There isn't enough information in the probability distribution to distinguish it from Happy Dance style problems.

Observation Counterfactuals

We can solve the problem in what *seems* like the right way by introducing a basic notion of counterfactual, which I'll write □→. This is supposed to represent "what the agent's code will do on different inputs". The idea is that if we have the policy of dancing when we see the money, M□→H is true *even in the world where we don't see any money*. So, even if dancing upon seeing money is a priori probable, conditioning on not doing so knocks out just as much probability mass from non-money worlds as from money worlds. However, if a counterfactual A□→B is true *and* A is true, then its consequent BB must also be true. So, conditioning on a policy does change the probability of taking actions in the expected way.

In Happy Dance, there is no correlation between M□→H and M; so, we can condition on M□→H and M□→¬H to decide which policy is better, and get the result we expect. In Newcomb's problem, on the other hand, there *is* a correlation between the policy

chosen and whether the pile of money appears, because Omega is checking what the agent's code does if it sees different inputs. This allows the decision theory to produce different answers in the different problems.

It's not clear where the beliefs about this correlation come from, so these counterfactuals are still *almost* as mysterious as explicitly giving conditional probabilities for everything given different policies. However, it does seem to say something nontrivial about the structure of reasoning.

Also, note that these counterfactuals are in the opposite direction from what we normally think about: rather than the counterfactual consequences of actions we didn't take, now we need to know the counterfactual actions we'd take under outcomes we didn't see!

Living in an Inadequate World

Follow-up to: Moloch's Toolbox (pt. 1, pt. 2)

Be warned: Trying to put together a background model like the one I sketched in the previous chapter is a pretty perilous undertaking, especially if you don't have a professional economist checking your work at every stage.

Suppose I offered the following much simpler explanation of how babies are dying inside the US healthcare system:

What if parents don't really care about their babies?

Maybe parents don't bond to their babies so swiftly? Maybe they don't really care *that* much about those voiceless pink blobs in the early days? Maybe this is one of those things that people think they're *supposed* to feel very strongly, and yet the emotion isn't actually there. Maybe parents just sort of inwardly shrug when their infants die, and only pretend to be sad about it. If they really cared, wouldn't they demand a system that didn't kill babies?

In our taxonomy, this would be a "decisionmaker is not beneficiary" explanation, with the parents and doctors being the decisionmakers, and the babies being the beneficiaries.

A much simpler hypothesis, isn't it?

When we try to do inadequacy analysis, there is such a thing as *wrong guesses* and *false cynicism.*

I'm sure there are some parents who don't bond to their babies all that intensely. I'm sure some of them lie to themselves about that. But in the early days when Omegaven was just plain illegal to sell across state lines, some parents would drive for hours, every month, to buy Omegaven from the Boston Children's Hospital to take back to their home state. I, for one, would call that an *extraordinary effort*. Those parents went far outside their routine, beyond what the System would demand of them, beyond what the world was set up to support them doing by default. Most people won't make an effort that far outside their usual habits even if their own personal lives are at stake.

If parents are letting their babies die of liver damage *because the parents don't care*, we should find few extraordinary efforts in these and other cases of baby-saving. This is an observational consequence we can check, and the observational check fails to support the theory.

For a fixed amount of inadequacy, there is only so much dysfunction that needs to be invoked to explain it. By the nature of inadequacy there will usually be more than one thing going wrong at a time… but even so, there's only a bounded amount of failure to be explained. Every possible dysfunction is *competing* against every other possible dysfunction to explain the observed data. Sloppy cynicism will usually be wrong, just

like your Facebook acquaintances who attribute civilizational dysfunctions to giant malevolent conspiracies.

If you're sloppy, then you're almost always going to find some way to conclude, "Oh, those physicists are just part of the broken academic system, what would they really know about the Higgs boson?" You will detect inadequacy every time you go looking for it, whether or not it's there. If you see the same vision wherever you look, that's the same as being blind.

i.

In most cases, you won't need to resort to complicated background analyses to figure out whether something is broken.

I mean, it's not like the only possible way one might notice that the US health care system is a vast, ill-conceived machine that is broken and also on fire is to understand microeconomics and predict *a priori* that aspects of this system design might promote inadequate equilibria. In real life, one notices the brokenness by reading economists who blog about the grinding gears and seas of flame, and listening to your friends sob about the screams coming from the ruins.

Then what good does it do to understand Moloch's toolbox? What's the point of the skill?

I suspect that for many people, the primary benefit of inadequacy analysis will be in undoing a mistake already made, where they disbelieve in inadequacy even when they're looking straight at it.

There are people who would simply never *try* to put up 130 light bulbs in their house—because if that worked, surely some good and diligent professional researcher would have already tried it. The medical system would have made it a standard treatment, right? The doctor would already know about it, right? And sure, sometimes people are stupid, but we're also people and we're also stupid so how could we amateurs possibly do better than current researchers on SAD, et cetera.

Often the most commonly applicable benefit from a fancy rational technique will be to cancel out fancy irrationality.[1] I expect that the most common benefit of inadequacy analysis will be to break a certain kind of blind trust—that is, trust arrived at by mental reasoning processes that are insensitive to whether you actually inhabit a universe that's worthy of that trust—and open people's eyes to the blatant brokenness of things that are easily *observed* to be broken. Understanding the background theory helps cancel out the elaborate arguments saying that you *can't* second-guess the European Central Bank even when it's straightforward to show how and why they're making a mistake.

Conversely, I've also watched some people plunge straight into problems that I'd guess were inexploitable, without doing the check, and then fail—usually falling prey to the Free Energy Fallacy, supposing that they can win just by doing better on the axis they care about. That subgroup might benefit, not from being told, "Shut up, you'll always fail, the answer is always no," but just from a reminder to *check* for signs of inexploitability.

It may be that some of those people will end up always saying, “I can think of at least one Moloch’s toolbox element in play, therefore this problem will be exploitable!” No humanly possible strictures of rationality can be strict enough to prevent a really determined person from shooting themselves in the foot. But it does help to be aware that the skill exists, before you start refining the skill.

Whether you’re trying to move past modesty or overcome the Free Energy Fallacy:

- Step one is to realize that here is a place to build an explicit domain theory—to *want* to understand the meta-principles of free energy, the principles of Moloch’s toolbox and the converse principles that imply real efficiency, and build up a model of how they apply to various parts of the world.
- Step two is to adjust your mind’s exploitability detectors until they’re not *always* answering, “You couldn’t possibly exploit this domain, foolish mortal,” or, “Why trust those hedge-fund managers to price stocks correctly when they have such poor incentives?”

And then you can move on to step three: the fine-tuning against reality.

ii.

In my past experience, I’ve both undershot and overshot the relative competence of doctors in the US medical system:

Anecdote 1: I once became very worried when my then-girlfriend got a headache and started seeing blobs of color, and when she drew the blobs they were left-right asymmetrical. I immediately started worrying about the asymmetry, thinking, “This is the kind of symptom I’d expect if someone had suffered damage to just one side of the brain.” Nobody at the emergency room seemed very concerned, and she waited for a couple of hours to be seen, when I could remember reading that strokes had to be treated within the first few hours (better yet, minutes) to save as much brain tissue as possible.

What she was really experiencing, of course, was her first migraine. And I expect that every nurse we talked to *knew that*, but only a doctor is allowed to make *diagnoses*, so they couldn’t legally *tell us*. I’d read all sorts of wonderful papers about exotic and illuminating forms of brain damage, but no papers about the *much more common* ailments that people in emergency rooms actually have. “Think horses, not zebras,” as the doctors say.

Anecdote 2: I once saw a dermatologist for a dandruff problem. He diagnosed me with eczema, and gave me some steroid cream to put on my head for when the eczema became especially severe. It didn’t cure the dandruff—but I’d seen a doctor, so I shrugged and concluded that there probably wasn’t much to be done, since I’d already tried and failed using the big guns of the Medical System.

Eight years later, when I was trying to compound a ketogenic meal replacement fluid I’d formulated in an attempt to lose weight, my dandruff seemed to get worse. So I checked whether online paleo blogs had anything to say about treating dandruff via diet. I learned that a lot of dandruff is caused by the *Candida* fungus (which I’d never heard of), and that *the fungus eats ketones*. So if switching to a ketogenic diet (or

drinking MCT oil, which gets turned into ketones) makes your dandruff worse, why, your dandruff is probably the *Candida* fungus. I looked up what kills *Candida*, found that I should use a shampoo containing ketoconazole, kept Googling, found a paper stating that 2% ketocanozole shampoo is an order of magnitude more effective than 1%, learned that only 1% ketocanozole shampoo was sold in the US, and ordered imported 2% Nizoral from Thailand via Amazon. Shortly thereafter, dandruff was no longer a significant issue for me and I could wear dark shirts without constantly checking my right shoulder for white specks. If my dermatologist knew anything about dandruff commonly being caused by a fungus, he never said a word.

From those two data points and others like them, I infer that medical competence—not medical absolute performance, but medical competence relative to what I can figure out by Googling—is high-variance. I shouldn't trust my doctor on significant questions without checking her diagnosis and treatment plan on the Internet, and I also shouldn't trust myself.

A lot of the times we put on our inadequacy-detecting goggles, we're deciding whether to trust some aspect of society to be more competent than ourselves. Part of the point of learning to think in economic terms about this question is to make it more natural to treat it as a technical question where specific lines of evidence can shift specific conclusions to varying degrees.

In particular, you don't need to be strictly better or worse than some part of society. The question isn't *about* ranking people, so you can be smarter in some ways and dumber in others. It can vary from minute to minute as the gods roll their dice.

By contrast, the modest viewpoint seems to me to have a very *social-status*-colored perspective on such things.

In the modest world, either you think you're better than doctors and all the civilization backing them, or you admit you're not as good and that you ought to defer to them.

If you don't defer to doctors, then you'll end up as one of those people who try feeding their children organic herbs to combat cancer; the outside view says that that's what happens to most non-doctors who dare to think they're *better* than doctors.

On the modest view, it's *not* that we hold up a thumb and eyeball the local competence level, based mostly on observation and a little on economic thinking; and then update on our observed relative performance; and sometimes say, "This varies a lot. I'll have to check each time."

Instead, every time you decide whether you think you can do better, *you are declaring what sort of person you are*.

For an example of what I mean here, consider writer Ozy Brennan's taxonomy:

> I think a formative moment for any rationalist—our "Uncle Ben shot by the mugger" moment, if you will—is the moment you go "holy shit, everyone in the world is fucking insane." […]
>
> Now, there are basically two ways you can respond to this.
>
> First, you can say "holy shit, everyone in the world is fucking insane. Therefore, if I adopt the radical new policy of not being fucking insane, I can pick up these giant

> piles of utility everyone is leaving on the ground, and then I win." […]
>
> This is the strategy of discovering a hot new stock tip, investing all your money, winning big, and retiring to Maui.
>
> Second, you can say "holy shit, everyone in the world is fucking insane. However, none of them seem to realize that they're insane. By extension, I am probably insane. I should take careful steps to minimize the damage I do." […]
>
> This is the strategy of discovering a hot new stock tip, realizing that most stock tips are bogus, and not going bankrupt.[2]

According to this sociological hypothesis, people can react to the discovery that "everyone in the world is insane" by adopting the Maui strategy, or they can react by adopting the not-going-bankrupt strategy.

(Note the inevitable comparison to financial markets—the one part of civilization that worked well enough to prompt an economist, Eugene Fama, to come up with the modern notion of efficiency.)

Brennan goes on to say that these two positions form a "dialectic," but that nonetheless, some kinds of people are clearly on the "becoming-sane side of things" while others are more on the "insanity-harm-reduction side of things."

But, speaking first to the basic dichotomy that's being proposed, the whole point of becoming sane is that your beliefs *shouldn't* reflect what sort of person you are. To the extent you're succeeding, at least, your beliefs should just reflect how the world is.

Good reasoners don't believe that there are goblins in their closets. The ultimate reason for this isn't that goblin-belief is archaic, outmoded, associated with people lost in fantasy worlds, too much like wishful thinking, et cetera. It's just that we opened up our closets and looked and we didn't see any goblins.

The goal is simply to be the sort of person who, in worlds with closet goblins, ends up believing in closet goblins, and in worlds without closet goblins, ends up disbelieving in closet goblins. Avoiding beliefs that sound archaic does relatively little to help you learn that there are goblins in a world where goblins exist, so it does relatively little to establish that there aren't goblins in a world where they don't exist. Examining particular empirical predictions of the goblin hypothesis, on the other hand, *does* provide strong evidence about what world you're in.

To reckon with the discovery that the world is mad, Brennan suggests that we consider the mix of humble and audacious "impulses in our soul" and try to strike the right balance. Perhaps we have some personality traits or biases that dispose us toward believing in goblins, and others that dispose us toward doubting them. On this framing, the heart of the issue is how we can resolve this inner conflict; the heart isn't any question about the behavioral tendencies or physiology of goblins.

This is a central disagreement I have with modest epistemology: modest people end up believing that they live in an inexploitable world *because they're trying to avoid acting like an arrogant kind of person*. Under modest epistemology, you're not supposed to adapt rapidly and without hesitation to the realities of the situation as you observe them, because that would mean trusting *yourself* to assess adequacy levels; but you can't trust yourself, because Dunning-Kruger, et cetera.

The alternative to modest epistemology isn't an *immodest* epistemology where you decide that you're higher status than doctors after all and conclude that you can now invent your own *de novo* medical treatments as a matter of course. The alternative is deciding for yourself whether to trust yourself more than a particular facet of your civilization at this particular time and place, checking the results whenever you can, and building up skill.

When it comes to medicine, I try to keep in mind that anyone whatsoever with more real-world medical experience may have me beat *cold solid* when it comes to any real-world problem. And then I go right on double-checking online to see if I believe what the doctor tells me about whether consuming too much medium-chain triglyceride oil could stress my liver.[3]

In my experience, people who don't viscerally understand Moloch's toolbox and the ubiquitously broken Nash equilibria of real life and how group insanity can arise from intelligent individuals responding to their own incentives tend to unconsciously translate *all* assertions about relative system competence into assertions about relative status. If you don't see systemic competence as rare, or don't see real-world systemic competence as driven by rare instances of correctly aligned incentives, all that's *left* is status. All good and bad output is just driven by good and bad individual people, and to suggest that you'll have better output is to assert that you're individually smarter than everyone else. (This is what status hierarchy feels like from the inside: to perform better is to *be* better.)

On a trip a couple of years ago to talk with the European existential risk community, which has internalized norms from modest epistemology to an even greater extent than the Bay Area community has, I ran into various people who asked questions like, "Why do you and your co-workers at MIRI think you can do better than academia?" (MIRI is the Machine Intelligence Research Institute, the organization I work at.)

I responded that we were a small research institute that sustains itself on individual donors, thereby sidestepping a set of standard organizational demands that collectively create bad incentives for the kind of research we're working on. I described how we had deliberately organized ourselves to steer clear of incentives that discourage long-term substantive research projects, to avoid academia's "publish or perish" dynamic, and more generally to navigate around the multiple frontiers of competitiveness where researchers have to spend all their energy competing along those dimensions to get into the best journals.

These are known failure modes that academics routinely complain about, so I wasn't saying anything novel or clever. The point I wanted to emphasize was that it's not enough to say that you want risky long-term research in the abstract; you have to accept that your people won't be at the competitive frontier for journal publications anymore.

The response I got back was something like a divide-by-zero error. Whenever I said "the nonprofit I work at has different incentives that look *prima facie* helpful for solving this set of technical problems," my claim appeared to get parsed as "the nonprofit I work at is *better* (higher status, more authoritative, etc.) than academia."

I think that the people I was talking with had already internalized the mathematical concept of Nash equilibria, but I don't think they were steeped in a no-free-energy microeconomic equilibrium view of all of society where *most of the time* systems end up dumber than the people in them due to multiple layers of terrible incentives, and

that this is normal and not at all a surprising state of affairs to suggest. And if you haven't practiced thinking about organizations' comparative advantages from that perspective long enough to make that lens *more cognitively available* than the status comparisons lens, then it makes sense that all talk of relative performance levels between you and doctors, or you and academia, or whatever, will be autoparsed by the easier, more native, more automatic status lens.

Because, come on, do you *really* think you're more authoritative / respectable / qualified / reputable / adept than your doctor about medicine? If you think *that*, won't you start consuming Vitamin C megadoses to treat cancer? And if you're *not* more authoritative / respectable / qualified / reputable / adept than your doctor, then how could you possibly do better by doing Internet research?

(Among most people I know, the relative status feeling frequently gets verbalized in English as "smarter," so if the above paragraph didn't make sense, try replacing the social-status placeholder "authoritative / respectable / etc." with "smarter.")

Again, a lot of the benefit of becoming fluent with this viewpoint is just in having a way of seeing "systems with not-all-that-great outputs," often observed extensively and directly, that can parse into *something* that isn't "Am I higher-status ('smarter,' 'better,' etc.) than the people in the system?"

iii.

I once encountered a case of (honest) misunderstanding from someone who thought that when I cited something as an example of civilizational inadequacy (or as I put it at the time, "People are crazy and the world is mad"), the thing I was trying to argue was that the Great Stagnation was just due to unimpressive / unqualified / low-status ("stupid") scientists.[4] He thought I thought that all we needed to do was take people in our social circle and have them go into biotech, or put scientists through a CFAR unit, and we'd see huge breakthroughs.[5]

"*What?*" I said.

(I was quite surprised.)

"I never said anything like that," I said, after recovering from the shock. "You can't lift a ten-pound weight with one pound of force!"

I went on to say that it's conceivable you could get faster-than-current results if CFAR's annual budget grew 20x, and then they spent four years iterating experimentally on techniques, and then a group of promising biotechnology grad students went through a year of CFAR training…[6]

So another way of thinking about the central question of civilizational inadequacy is that we're trying to assess the *quantity of effort* required to achieve a given level of outperformance. Not "Can it be done?" but "How much work?"

This brings me to the single most obvious notion that correct contrarians grasp, and that people who have vastly overestimated their own competence don't realize: It

takes *far* less work to identify the correct expert in a pre-existing dispute between experts, than to make an *original contribution* to any field that is remotely healthy.

I did not work out myself what would be a better policy for the Bank of Japan. I believed the arguments of Scott Sumner, who is not literally mainstream (yet), but whose position is shared by many other economists. I sided with a particular band of contrarian expert economists, based on my attempt to parse the object-level arguments, observing from the sidelines for a while to see who was right about near-term predictions and picking up on what previous experience suggested were strong cues of correct contrarianism.[7]

And so I ended up thinking that I knew better than the Bank of Japan. On the modest view, that's just about as immodest as thinking you can personally advance the state of the art, since who says I ought to be smarter than the Bank of Japan at picking good experts to trust, et cetera?

But in real life, inside a civilization that is often tremendously broken on a systemic level, finding a contrarian expert seeming to shine against an untrustworthy background is *nowhere remotely near* as difficult as becoming that expert yourself. It's the difference between picking which of four runners is most likely to win a fifty-kilometer race, and winning a fifty-kilometer race yourself.

Distinguishing a correct contrarian isn't easy in absolute terms. You are still trying to be better than the mainstream in deciding who to trust.[8] For many people, yes, an attempt to identify contrarian experts ends with them trusting faith healers over traditional medicine. But it's still in the range of things that amateurs can do with a reasonable effort, if they've picked up on unusually good epistemology from one source or another.

We live in a sufficiently poorly-functioning world that there are many visibly correct contrarians whose ideas are not yet being implemented in the mainstream, where the authorities who allegedly judge between experts are making errors that appear to me trivial. (And again, by "errors," I mean that these authorities are endorsing factually wrong answers or dominated policies—*not* that they're passing up easy rewards given their incentives.)

In a world like that, you can often know things that the average authority doesn't know… but *not* because you figured it out yourself, in almost every case.

iv.

Going beyond picking the right horse in the race and becoming a horse yourself, inventing your own new personal solution to a civilizational problem, requires a much greater investment of effort.

I did make up my own decision theory—not from a *tabula rasa*, but still to my own recipe. But events like that should be *rare* in a given person's life. Logical counterfactuals in decision theory are one of my *few* major contributions to an existing academic field, and my early thoughts on this topic were quickly improved on by others.[9] And that was a significant life event, not the sort of thing I believe I've done every month.

Above all, reaching the true frontier requires *picking your battles*.

Computer security professionals don't attack systems by picking one particular function and saying, "Now I shall find a way to exploit these exact 20 lines of code!" Most lines of code in a system don't provide exploits no matter how hard you look at them. In a large enough system, there are rare lines of code that are exceptions to this general rule, and sometimes you can be the first to find them. But if we think about a random section of code, the base rate of exploitability is extremely low—except in *really, really bad code* that nobody looked at from a security standpoint in the first place.

Thinking that you've searched a large system and found one new exploit is one thing. Thinking that you can exploit arbitrary lines of code is quite another.

No matter how broken academia is, no one can improve on arbitrary parts of the modern academic edifice. My own base frequency for seeing scholarship that I think I can improve upon is "almost never," outside of some academic subfields dealing with the equivalent of "unusually bad code." But don't expect bad code to be guarding vaults of gleaming gold in a form that other people value, except with a very low base rate. There do tend to be real locks on the energy-containing vaults not already emptied… *almost* (but not quite) all of the time.

Similarly, you do not generate a good startup idea by taking some random activity, and then talking yourself into believing you can do it better than existing companies. Even where the current way of doing things seems bad, and even when you really do know a better way, 99 times out of 100 you will not be able to make money by knowing better. If somebody else makes money on a solution to that particular problem, they'll do it using rare resources or skills that you don't have—including the skill of being super-charismatic and getting tons of venture capital to do it.

To believe you have a good startup idea is to say, "Unlike the typical 99 cases, in this particular anomalous and unusual case, I think I *can* make a profit by knowing a better way."

The anomaly doesn't have to be some super-unusual skill possessed by you alone in all the world. That would be a question that always returned "No," a blind set of goggles. Having an unusually good idea might work well enough to be worth trying, if you think you can standardly solve the other standard startup problems. I'm merely emphasizing that to find a rare startup idea that is *exploitable* in dollars, you will have to *scan and keep scanning,* not pursue the first "*X* is broken and maybe I can fix it!" thought that pops into your head.

To win, choose winnable battles; await the rare *anomalous* case of, "Oh wait, that could work."

v.

In 2014, I experimentally put together my own ketogenic meal replacement drink via several weeks of research, plus months of empirical tweaking, to see if it could help me with long-term weight normalization.

In that case, I did not get to pick my battleground.

And yet even so, I still tried to design my own recipe. Why? It seems I must have thought I could do better than the best ketogenic liquid-food recipes that had ever before been tried, as of 2014. Why would I believe I could do the best of anyone who's yet tried, when I couldn't pick my battle?

Well, because I looked up previous ketogenic Soylent recipes, and they used standard multivitamin powders containing, e.g., way too much manganese and the wrong form of selenium. (You get all the manganese you need from ordinary drinking water, if it hasn't been distilled or bottled. Excess amounts may be *neurotoxic*. One of the leading hypotheses for why multivitamins aren't found to produce net health improvement, despite having many individual components found to be helpful, is that multivitamins contain 100% of the US RDA of manganese. Similarly, if a multivitamin includes sodium selenite instead of, e.g., se-methyl-selenocysteine, it's the equivalent of handing you a lump of charcoal and saying, "You're a carbon-based lifeform; this has carbon in it, right?")

Just for the sake of grim amusement, I also looked up my civilization's medically standard ketogenic dietary options—e.g., for epileptic children. As expected, they were far worse than the amateur Soylent-inspired recipes. They didn't even contain medium-chain triglycerides, which your liver turns directly into ketones. (MCT is academically recommended, though not commercially standard, as the basis for maintaining ketosis in epileptic children.) Instead the retail dietary options for epileptic children involved mostly soybean oil, of which it has been said, "Why not just shoot them?"

Even when we can't pick our battleground, sometimes the most advanced weapon on offer turns out to be a broken stick and it's worth the time to carve a handaxe.

… But even then, I didn't try to synthesize my own dietary *theory* from scratch. There is nothing I believe about how human metabolism works that's unique or original to me. Not a single element of my homemade Ketosoylent was based on my personal, private theory of how *any* of the micronutrients worked. Who am I to think I understand Vitamin D3 better than everyone else in the world?

The Ketosoylent didn't work for long-term weight normalization, alas—the same result as all other replicated experiments on trying to long-term-normalize weight via putting different things inside your mouth. (The Shangri-La Diet I mentioned at the start of this book didn't work for me either.)

So it goes. I mention the Ketosoylent because it's the most complicated thing I've tried to do *without* tons of experience in a domain and *without* being able to pick my battles.

In the simpler and happier case of treating Brienne's Seasonal Affective Disorder, I again didn't get to pick the battleground; but SAD has received far less scientific attention to date than obesity. And success there again didn't involve coming up with an amazing new model of SAD. It's not weird and private knowledge that sufficiently bright light might cure SAD. The Sun is known to work almost all the time.

So a realistic lifetime of trying to adapt yourself to a broken civilization looks like:

- 0–2 lifetime instances of answering "Yes" to "Can I substantially improve on my civilization's current knowledge *if I put years into the attempt?*" A few people, but not many, will answer "Yes" to enough instances of this question to count on

the fingers of both hands. Moving on to your toes indicates that you are a crackpot.

- Once per year or thereabouts, an answer of “Yes” to “Can I generate a synthesis of existing correct contrarianism which will beat my current civilization’s next-best alternative, for just myself (i.e., without trying to solve the further problems of widespread adoption), after a few weeks’ research and a bunch of testing and occasionally asking for help?” (See my experiments with ketogenic diets and SAD treatment; also what you would do to generate or judge a startup idea that wasn’t based on a hard science problem.)

- *Many* cases of trying to pick a previously existing side in a running dispute between experts, if you think that you can follow the object-level arguments reasonably well and there are strong meta-level cues that you can identify.

The accumulation of many judgments of the latter kind is where you get the fuel for many small day-to-day decisions (e.g., about what to eat), and much of your ability to do larger things (like solving a medical problem after going through the medical system has proved fruitless, or executing well on a startup).

vi.

A few final pieces of advice on everyday thinking about inadequacy:

When it comes to estimating the competence of some aspect of civilization, especially relative to your own competence, try to update hard on your experiences of failure and success. One data point is a hell of a lot better than zero data points.

Worrying about how one data point is “just an anecdote” can make sense if you’ve already collected thirty data points. On the other hand, when you previously just had a lot of prior reasoning, or you were previously trying to generalize from other people’s not-quite-similar experiences, and then you collide directly with reality for the first time, one data point is *huge.*

If you do accidentally update too far, you can always re-update later when you have more data points. So update hard on each occasion, and take care not to flush any new observation down the toilet.

Oh, and bet. Bet on everything. Bet real money. It helps a lot with learning.

I once bet $25 at even odds against the eventual discovery of the Higgs boson—after 90% of the possible mass range had been experimentally eliminated, because I had the impression from reading diatribes against string theory that modern theoretical physics might not be solid enough to predict a qualitatively new kind of particle with prior odds greater than 9:1.

When the Higgs boson was discovered inside the remaining 10% interval of possible energies, I said, “Gosh, I guess they *can* predict that sort of thing with prior probability greater than 90%,” updated strongly in favor of the credibility of things like dark matter and dark energy, and then didn’t make any more bets like that.

I made a mistake; and I bet on it. This let me *experience* the mistake in a way that helped me better learn from it. When you're thinking about large, messy phenomena like "the adequacy of human civilization at understanding nutrition," it's easy to get caught up in plausible-sounding stories and never quite get around to running the experiment. Run experiments; place bets; say *oops*. Anything less is an act of self-sabotage.

Next: **Blind Empiricism**.

The full book will be available November 16th. You can go to equilibriabook.com to pre-order the book, or sign up for notifications about new chapters and other developments.

---

1. As an example, relatively few people in the world need well-developed skills at cognitive reductionism for the purpose of disassembling aspects of nature. The reason why *anyone else* needs to learn cognitive reductionism—the reason it's this big public epistemic hygiene issue—is that there are a lot of damaging supernatural beliefs that cognitive reductionism helps counter. ↩

2. Brennan, "The World Is Mad."

   When I ran a draft of this chapter by Brennan, they said that they basically agree with what I'm saying here, but are thinking about these issues using a different conceptual framework. ↩

3. Answer: this is the opposite of standard theory; she was probably confusing MCT with other forms of saturated fat. ↩

4. The Great Stagnation is economist Tyler Cowen's hypothesis that declining rates of innovation since the 1970s (excluding information technology, for the most part) have resulted in relative economic stagnation in the developed world. ↩

5. CFAR, the Center for Applied Rationality, is a nonprofit that applies ideas from cognitive science to everyday problem-solving and decision-making, running workshops for people who want to get better at solving big global problems. MIRI and CFAR are frequent collaborators, and share office space; the organization's original concept came from MIRI's work on rationality. ↩

6. See also Weinersmith's Law: "*No problem is too hard. Many problems are too fast.*" ↩

7. E.g., the cry of "Stop ignoring your own carefully gathered experimental evidence, damn it!" ↩

8. Though, to be clear, the mainstream isn't *actually* deciding who to trust. It's picking winners by some other criterion that on a good day is not totally uncorrelated with trustworthiness. ↩

9. In particular, Wei Dai came up with updatelessness, yielding the earliest version of what's now called functional decision theory. See Soares and Levinstein's "Cheating Death in Damascus" for a description. ↩

Mapping Another's Universe

Let's say you are reading a book about people with magical abilities. As a young child, their abilities manifest spontaneously (accidentally breaking something with their mind, flying instead of falling, etc). Then, they are taken away off somewhere to learn to hone their skills. There may be rules to what magic is and is not possible, how magic is done, and what different students can learn. In this book, the main character has the ability to manipulate electricity. They can use this to control anything that operates with electricity remotely, like by turning lights on and off, or they can just shock people. If it is established in the book that no person has more than one ability, you may be surprised if, later on, the main character starts to manipulate water or read minds. This changes the rules of the story.

In the above paragraph, I painted a picture of a universe, giving you rules about how it worked. If I asked you to write a story within the universe, many of you might take care to make sure the story abides by the rules I have provided. This ability to create a universe in your mind is a skill not everyone has, but it is useful beyond just reading stories and writing fanfiction. This same skill can be used to understand people, and even ourselves, better.

To tie this to another example, let's say you are speaking with someone who you just met. This person is an elementary school teacher. Right away, you know quite a few things about this person (they work with kids, get summers off, probably teach during the day and maybe grade papers and plan lessons after school). In the same way that you did with the book, you can use this information to give more detail to this person's universe. Now maybe the person tells you that they also run a summer camp. If you previously assumed that being a teacher meant they had summers off, now you can update the information you have on this person to include that they have a second job they work over the summer. You now have a slightly more detailed picture of what their universe is like.

Though the example above is a relatively simple one, there are lots of ways to gain information that you can use to add detail to someone's universe. Often times we don't even realize we are doing it; assumptions made based on a person's appearance or social media profile aren't always conscious. There are even small things people say and do that we miss, but could otherwise be useful information in understanding the person better. If someone tells you a movie they saw was "too scary", it may not be a fact about the movie, but about their dislike for horror movies, or fears surrounding the topic of the movie.

The perspective we use to analyze other universes is also important, because it determines what we notice and what we don't. When reading books, we may not think much of things that are normal for us, like a 6-year-old going to school for the first time, but notice things that are not true in our universe, like super powers. The more ways a fictional universe is different from our own, the harder it is to keep track of. Imagine how much easier it might be to follow a book about super heroes, rather than a book about aliens and monsters with magic and futuristic weapons! This is part of why we enjoy spending time with and talking with people who think similarly to us. If a friend's universe is similar to yours, it takes less effort to understand who they are and why they do what they do.

If you meet someone whose universe is too different from yours, you may find them hard to relate to. Part of this is because, by default, we use our own universe to look at other universes. A young child who doesn't know the meaning of divorce might have a hard time understanding the experiences of their friend, who spends half their time with Dad and half their time with Mom, but never together. The child may get frustrated if that friend leaves a book they borrowed at Dad's house, then didn't have it when they spent time together after school at Mom's. It can be so easy to judge other people without thinking about whether what happened makes sense in their universe if it doesn't make sense in yours. The child may assume that the friend didn't want to give the book back or was trying to be mean, rather than realize how easy it is to forget something at one house when you have two. Even adults do this all the time, assuming malice or stupidity when we can't understand another's actions. Taking care to understand another person's universe can help prevent this from happening, sometimes drawing attention to parts of your own universe that you take for granted in the process.

Security Mindset and the Logistic Success Curve

Follow-up to: Security Mindset and Ordinary Paranoia

(*Two days later, Amber returns with another question.*)

AMBER: Uh, say, Coral. How important is security mindset when you're building a whole new kind of system—say, one subject to potentially adverse optimization pressures, where you want it to have some sort of robustness property?

CORAL: How novel is the system?

AMBER: Very novel.

CORAL: Novel enough that you'd have to invent your own new best practices instead of looking them up?

AMBER: Right.

CORAL: That's serious business. If you're building a very simple Internet-connected system, maybe a smart ordinary paranoid could look up how we usually guard against adversaries, use as much off-the-shelf software as possible that was checked over by real security professionals, and not do too horribly. But if you're doing something qualitatively new and complicated that has to be robust against adverse optimization, well... mostly I'd think you were operating in almost impossibly dangerous territory, and I'd advise you to figure out what to do after your first try failed. But if you wanted to actually succeed, ordinary paranoia absolutely would not do it.

AMBER: In other words, projects to build novel mission-critical systems ought to have advisors with the full security mindset, so that the advisor can say what the system builders really need to do to ensure security.

CORAL: (*laughs sadly*) No.

AMBER: No?

CORAL: Let's say for the sake of concreteness that you want to build a new kind of secure operating system. That is *not* the sort of thing you can do by attaching one advisor with security mindset, who has limited political capital to use to try to argue people into doing things. “Building a house when you're only allowed to touch the bricks using tweezers” comes to mind as a metaphor. You're going to need experienced security professionals working full-time with high authority. Three of them, one of whom is a cofounder. Although even then, we might still be operating in the territory of Paul Graham's Design Paradox.

AMBER: Design Paradox? What's that?

CORAL: Paul Graham's Design Paradox is that people who have good taste in UIs can tell when other people are designing good UIs, but most CEOs of big companies lack the good taste to tell who else has good taste. And that's why big companies can't just hire other people as talented as Steve Jobs to build nice things for them, even though Steve Jobs certainly wasn't the best possible designer on the planet. Apple existed because of a lucky history where Steve Jobs ended up in charge. There's no way for Samsung to hire somebody else with equal talents, because Samsung would just end up with some guy in a suit who was good at pretending to be Steve Jobs in front of a CEO who couldn't tell the difference.

Similarly, people with security mindset can notice when other people lack it, but I'd worry that an ordinary paranoid would have a hard time telling the difference, which would make it hard for them to hire a truly competent advisor. And of course lots of the people in the larger social system behind technology projects lack even the ordinary paranoia that many good programmers possess, and they just end up with empty suits talking a lot about “risk” and “safety”. In other words, if we're talking about something as hard as building a secure operating system, and your project hasn't started up *already* headed up by someone with the full security mindset, you are in trouble. Where by “in trouble” I mean “totally, irretrievably doomed”.

AMBER: Look, uh, there's a certain project I'm invested in which has raised a hundred million dollars to create merchant drones.

CORAL: Merchant drones?

AMBER: So there are a lot of countries that have poor market infrastructure, and the idea is, we're going to make drones that fly around buying and selling things, and they'll use machine learning to figure out what prices to pay and so on. We're not just in it for the money; we think it could be a huge economic boost to those countries, really help them move forwards.

CORAL: Dear God. Okay. There are exactly two things your company is about: system security, and regulatory compliance. Well, and also marketing, but that doesn't count because every company is about marketing. It would be a severe error to imagine that your company is about anything else, such as drone hardware or machine learning.

AMBER: Well, the sentiment inside the company is that the time to begin thinking about legalities and security will be after we've proven we can build a prototype and have at least a small pilot market in progress. I mean, until we know how people are using the system and how the software ends up working, it's hard to see how we could do any productive thinking about security or compliance that wouldn't just be pure speculation.

CORAL: Ha! Ha, hahaha… oh my god you're not joking.

AMBER: What?

CORAL: Please tell me that what you actually mean is that you have a security and regulatory roadmap which calls for you to do some of your work later, but clearly lays out what work needs to be done, when you are to start doing it, and when each milestone needs to be complete. Surely you don't *literally* mean that you *intend to start thinking about it* later?

AMBER: A lot of times at lunch we talk about how annoying it is that we'll have to deal with regulations and how much better it would be if governments were more

libertarian. That counts as thinking about it, right?

CORAL: Oh my god.

AMBER: I don't see how we could have a security plan when we don't know exactly what we'll be securing. Wouldn't the plan just turn out to be wrong?

CORAL: All business plans for startups turn out to be wrong, but you still need them—and not just as works of fiction. They represent the written form of your current beliefs about your key assumptions. Writing down your business plan checks whether your current beliefs can possibly be coherent, and suggests which critical beliefs to test first, and which results should set off alarms, and when you are falling behind key survival thresholds. The idea isn't that you stick to the business plan; it's that having a business plan (a) checks that it seems possible to succeed in any way whatsoever, and (b) tells you when one of your beliefs is being falsified so you can explicitly change the plan and adapt. Having a written plan that you intend to rapidly revise in the face of new information is one thing. *NOT HAVING A PLAN* is *another*.

AMBER: The thing is, I *am* a little worried that the head of the project, Mr. Topaz, isn't concerned enough about the possibility of somebody fooling the drones into giving out money when they shouldn't. I mean, I've tried to raise that concern, but he says that of course we're not going to program the drones to give out money to just anyone. Can you maybe give him a few tips? For when it comes time to start thinking about security, I mean.

CORAL: Oh. Oh, my dear, sweet summer child, I'm sorry. There's nothing I can do for you.

AMBER: Huh? But you haven't even looked at our beautiful business model!

CORAL: I thought maybe your company merely had a hopeless case of underestimated difficulties and misplaced priorities. But now it sounds like your leader is not even using ordinary paranoia, and reacts with skepticism to it. Calling a case like that “hopeless” would be an understatement.

AMBER: But a security failure would be very bad for the countries we're trying to help! They need *secure* merchant drones!

CORAL: Then they will need drones built by some project that is not led by Mr. Topaz.

AMBER: But that seems very hard to arrange!

CORAL: ...I don't understand what you are saying that is supposed to contradict anything I am saying.

AMBER: Look, aren't you judging Mr. Topaz a little too quickly? Seriously.

CORAL: I haven't met him, so it's possible you misrepresented him to me. But if you've accurately represented his attitude? Then, yes, I did judge quickly, but it's a hell of a good guess. Security mindset is already rare on priors. “I don't plan to make my drones give away money to random people” means he's imagining how his system could work as he intends, instead of imagining how it might not work as he intends. If somebody doesn't even exhibit ordinary paranoia, spontaneously on their own cognizance without external prompting, then they cannot do security, period. Reacting

indignantly to the suggestion that something might go wrong is even beyond that level of hopelessness, but the base level was hopeless enough already.

AMBER: Look... can you just go to Mr. Topaz and try to tell him what he needs to do to add some security onto his drones? Just try? Because it's super important.

CORAL: I could try, yes. I can't succeed, but I could try.

AMBER: Oh, but please be careful to not be harsh with him. Don't put the focus on what he's doing wrong—and try to make it clear that these problems aren't *too* serious. He's been put off by the media alarmism surrounding apocalyptic scenarios with armies of evil drones filling the sky, and it took me some trouble to convince him that I wasn't just another alarmist full of fanciful catastrophe scenarios of drones defying their own programming.

CORAL: ...

AMBER: And maybe try to keep your opening conversation away from what might sound like crazy edge cases, like somebody forgetting to check the end of a buffer and an adversary throwing in a huge string of characters that overwrite the end of the stack with a return address that jumps to a section of code somewhere else in the system that does something the adversary wants. I mean, you've convinced me that these far-fetched scenarios are worth worrying about, if only because they might be canaries in the coal mine for more realistic failure modes. But Mr. Topaz thinks that's all a bit silly, and I don't think you should open by trying to explain to him on a meta level why it isn't. He'd probably think you were being condescending, telling him how to think. Especially when you're just an operating-systems guy and you have no experience building drones and seeing what actually makes them crash. I mean, that's what I think he'd say to you.

CORAL: ...

AMBER: Also, start with the cheaper interventions when you're giving advice. I don't think Mr. Topaz is going to react well if you tell him that he needs to start all over in another programming language, or establish a review board for all code changes, or whatever. He's worried about competitors reaching the market first, so he doesn't want to do anything that will slow him down.

CORAL: ...

AMBER: Uh, Coral?

CORAL: ... on his novel project, entering new territory, doing things not exactly like what has been done before, carrying out novel mission-critical subtasks for which there are no standardized best security practices, nor any known understanding of what makes the system robust or not-robust.

AMBER: Right!

CORAL: And Mr. Topaz himself does not seem much terrified of this terrifying task before him.

AMBER: Well, he's worried about somebody else making merchant drones first and misusing this key economic infrastructure for bad purposes. That's the same basic thing, right? Like, it demonstrates that he can worry about things?

CORAL: It is utterly different. Monkeys who can be afraid of other monkeys getting to the bananas first are far, far more common than monkeys who worry about whether the bananas will exhibit weird system behaviors in the face of adverse optimization.

AMBER: Oh.

CORAL: I'm afraid it is only slightly more probable that Mr. Topaz will oversee the creation of robust software than that the Moon will spontaneously transform into organically farmed goat cheese.

AMBER: I think you're being too harsh on him. I've met Mr. Topaz, and he seemed pretty bright to me.

CORAL: Again, assuming you're representing him accurately, Mr. Topaz seems to lack what I called ordinary paranoia. If he does have that ability as a cognitive capacity, which many bright programmers do, then he obviously doesn't feel passionate about applying that paranoia to his drone project along key dimensions. It also sounds like Mr. Topaz doesn't realize there's a skill that he is missing, and would be insulted by the suggestion. I am put in mind of the story of the farmer who was asked by a passing driver for directions to get to Point B, to which the farmer replied, “If I was trying to get to Point B, I sure wouldn't start from here.”

AMBER: Mr. Topaz has made some significant advances in drone technology, so he can't be stupid, right?

CORAL: "Security mindset" seems to be a distinct cognitive talent from *g* factor or even programming ability. In fact, there doesn’t seem to be a level of human genius that even guarantees you’ll be skilled at ordinary paranoia. Which does make some security professionals feel a bit weird, myself included—the same way a lot of programmers have trouble understanding why not everyone can learn to program. But it seems to be an observational fact that both ordinary paranoia and security mindset are things that can decouple from *g* factor and programming ability—and if this were not the case, the Internet would be far more secure than it is.

AMBER: Do you think it would help if we talked to the other VCs funding this project and got them to ask Mr. Topaz to appoint a Special Advisor on Robustness reporting directly to the CTO? That sounds politically difficult to me, but it's possible we could swing it. Once the press started speculating about drones going rogue and maybe aggregating into larger Voltron-like robots that could acquire laser eyes, Mr. Topaz did tell the VCs that he was very concerned about the ethics of drone safety and that he'd had many long conversations about it over lunch hours.

CORAL: I'm venturing slightly outside my own expertise here, which isn't corporate politics per se. But on a project like this one that's trying to enter novel territory, I'd guess the person with security mindset needs at least cofounder status, and must be personally trusted by any cofounders who don't have the skill. It can't be an outsider who was brought in by VCs, who is operating on limited political capital and needs to win an argument every time she wants to not have all the services conveniently turned on by default. I suspect you just have the wrong person in charge of this startup, and that this problem is not repairable.

AMBER: Please don't just give up! Even if things are as bad as you say, just increasing our project's probability of being secure from 0% to 10% would be very valuable in expectation to all those people in other countries who need merchant drones.

CORAL: ...look, at some point in life we have to try to triage our efforts and give up on what can't be salvaged. There's often a logistic curve for success probabilities, you know? The distances are measured in multiplicative odds, not additive percentage points. You can't take a project like this and assume that by putting in some more hard work, you can increase the absolute chance of success by 10%. More like, the odds of this project's failure versus success start out as 1,000,000:1, and if we're very polite and navigate around Mr. Topaz's sense that he is higher-status than us and manage to explain a few tips to him without ever sounding like we think we know something he doesn't, we can quintuple his chances of success and send the odds to 200,000:1. Which is to say that in the world of percentage points, the odds go from 0.0% to 0.0%. That's one way to look at the "law of continued failure".

If you had the kind of project where the fundamentals implied, say, a 15% chance of success, you'd then be on the right part of the logistic curve, and in *that* case it could make a lot of sense to hunt for ways to bump that up to a 30% or 80% chance.

AMBER: Look, I'm worried that it will really be very bad if Mr. Topaz reaches the market first with insecure drones. Like, I think that merchant drones could be very beneficial to countries without much existing market backbone, and if there's a grand failure—especially if some of the would-be customers have their money or items stolen—then it could poison the potential market for years. It will be terrible! Really, genuinely terrible!

CORAL: Wow. That sure does sound like an unpleasant scenario to have wedged yourself into.

AMBER: But what do we do now?

CORAL: Damned if I know. I do suspect you're screwed so long as you can only win if somebody like Mr. Topaz creates a robust system. I guess you could try to have some other drone project come into existence, headed up by somebody that, say, Bruce Schneier assures everyone is unusually good at security-mindset thinking and hence can hire people like me and listen to all the harsh things we have to say. Though I have to admit, the part where you think it's drastically important that you beat an insecure system to market with a secure system—well, that sounds positively nightmarish. You're going to need a lot more resources than Mr. Topaz has, or some other kind of very major advantage. Security takes time.

AMBER: Is it really that hard to add security to the drone system?

CORAL: You keep talking about "adding" security. System robustness isn't the kind of property you can bolt onto software as an afterthought.

AMBER: I guess I'm having trouble seeing why it's so much more expensive. Like, if somebody foolishly builds an OS that gives access to just anyone, you could instead put a password lock on it, using your clever system where the OS keeps the hashes of the passwords instead of the passwords. You just spend a couple of days rewriting all the services exposed to the Internet to ask for passwords before granting access. And then the OS has security on it! Right?

CORAL: NO. Everything inside your system that is potentially subject to adverse selection in its probability of weird behavior is a liability! Everything exposed to an attacker, and everything those subsystems interact with, and everything *those* parts interact with! You have to build *all* of it robustly! If you want to build a secure OS you need a whole special project that is "building a secure operating system instead of an

insecure operating system”. And you also need to restrict the scope of your ambitions, and not do everything you want to do, and obey other commandments that will feel like big unpleasant sacrifices to somebody who doesn't have the full security mindset. OpenBSD can’t do a tenth of what Ubuntu does. They can't afford to! It would be too large of an attack surface! They can't review that much code using the special process that they use to develop secure software! They can't hold that many assumptions in their minds!

AMBER: Does that effort *have* to take a significant amount of extra time? Are you sure it can't just be done in a couple more weeks if we hurry?

CORAL: YES. Given that this is a novel project entering new territory, expect it to take *at least* two years more time, or 50% more development time—whichever is less—compared to a security-incautious project that otherwise has identical tools, insights, people, and resources. And that is a very, very optimistic lower bound.

AMBER: This story seems to be heading in a worrying direction.

CORAL: Well, I'm sorry, but creating robust systems takes longer than creating non-robust systems even in cases where it would be really, extraordinarily bad if creating robust systems took longer than creating non-robust systems.

AMBER: Couldn't it be the case that, like, projects which are implementing good security practices do everything so much cleaner and better that they can come to market faster than any insecure competitors could?

CORAL: … I honestly have trouble seeing why you’re privileging that hypothesis for consideration. Robustness involves assurance processes that take additional time. OpenBSD does not go through lines of code faster than Ubuntu.

But more importantly, if everyone has access to the same tools and insights and resources, then an unusually fast method of doing something cautiously can always be degenerated into an even faster method of doing the thing incautiously. There is not now, nor will there ever be, a programming language in which it is the least bit difficult to write bad programs. There is not now, nor will there ever be, a methodology that makes writing insecure software inherently slower than writing secure software. Any security professional who heard about your bright hopes would just laugh. Ask them too if you don't believe me.

AMBER: But shouldn't engineers who aren't cautious just be unable to make software at all, because of ordinary bugs?

CORAL: I am afraid that it is both possible, and *extremely* common in practice, for people to fix all the bugs that are crashing their systems in ordinary testing today, using methodologies that are indeed adequate to fixing ordinary bugs that show up often enough to afflict a significant fraction of users, and then ship the product. They get everything working today, and they don't feel like they have the slack to delay any longer than that before shipping because the product is already behind schedule. They don't hire exceptional people to do ten times as much work in order to prevent the product from having holes that only show up under adverse optimization pressure, that somebody else finds first and that they learn about after it's too late.

It's not even the wrong decision, for products that aren't connected to the Internet, don't have enough users for one to go rogue, don't handle money, don't contain any valuable data, and don't do anything that could injure people if something goes

wrong. If your software doesn't destroy anything important when it explodes, it's probably a better use of limited resources to plan on fixing bugs as they show up.

… Of course, you need some amount of security mindset to realize which software *can* in fact destroy the company if it silently corrupts data and nobody notices this until a month later. I don't suppose it's the case that your drones only carry a limited amount of the full corporate budget in cash over the course of a day, and you always have more than enough money to reimburse all the customers if all items in transit over a day were lost, taking into account that the drones might make many more purchases or sales than usual? And that the systems are generating internal paper receipts that are clearly shown to the customer and non-electronically reconciled once per day, thereby enabling you to notice a problem before it's too late?

AMBER: Nope!

CORAL: Then as you say, it would be better for the world if your company didn't exist and wasn't about to charge into this new territory and poison it with a spectacular screwup.

AMBER: If I believed that… well, Mr. Topaz certainly isn't going to stop his project or let somebody else take over. It seems the logical implication of what you say you believe is that I should try to persuade the venture capitalists I know to launch a safer drone project with even more funding.

CORAL: Uh, I'm sorry to be blunt about this, but I'm not sure *you* have a high enough level of security mindset to identify an executive who's sufficiently better than you at it. Trying to get enough of a resource advantage to beat the insecure product to market is only half of your problem in launching a competing project. The other half of your problem is surpassing the prior rarity of people with truly deep security mindset, and getting somebody like that in charge and fully committed. Or at least get them in as a highly trusted, fully committed cofounder who isn't on a short budget of political capital. I'll say it again: an advisor appointed by VCs isn't nearly enough for a project like yours. Even if the advisor is a genuinely good security professional—

AMBER: This all seems like an unreasonably difficult requirement! Can't you back down on it a little?

CORAL: —the person in charge will probably try to bargain down reality, as represented by the unwelcome voice of the security professional, who won't have enough social capital to badger them into “unreasonable” measures. Which means you fail on full automatic.

AMBER: … Then what am I to do?

CORAL: I don't know, actually. But there's no point in launching another drone project with even more funding, if it just ends up with another Mr. Topaz put in charge. Which, by default, is exactly what your venture capitalist friends are going to do. Then you've just set an even higher competitive bar for anyone actually trying to be first to market with a secure solution, may God have mercy on their souls.

Besides, if Mr. Topaz thinks he has a competitor breathing down his neck and rushes his product to market, his chance of creating a secure system could drop by a factor of ten and go all the way from 0.0% to 0.0%.

AMBER: Surely my VC friends have faced this kind of problem before and know how to identify and hire executives who can do security well?

CORAL: … If one of your VC friends is Paul Graham, then maybe yes. But in the average case, *NO*.

If average VCs always made sure that projects which needed security had a founder or cofounder with strong security mindset—if they had the *ability* to do that *even in cases where they decided they wanted to*—the Internet would again look like a very different place. By default, your VC friends will be fooled by somebody who looks very sober and talks a lot about how terribly concerned he is with cybersecurity and how the system is going to be ultra-secure and reject over nine thousand common passwords, including the thirty-six passwords listed on this slide here, and the VCs will ooh and ah over it, especially as one of them realizes that their own password is on the slide. *That* project leader is absolutely not going to want to hear from me—even less so than Mr. Topaz. To him, I'm a political threat who might damage his line of patter to the VCs.

AMBER: I have trouble believing all these smart people are really that stupid.

CORAL: You're compressing your innate sense of social status and your estimated level of how good particular groups are at this particular ability into a single dimension. That is not a good idea.

AMBER: I'm not saying that I think everyone with high status already knows the deep security skill. I'm just having trouble believing that they can't learn it quickly once told, or could be stuck not being able to identify good advisors who have it. That would mean they couldn't know something you know, something that seems important, and that just… feels *off* to me, somehow. Like, there are all these successful and important people out there, and you're saying you're *better* than them, even with all their influence, their skills, their resources—

CORAL: Look, you don't have to take my word for it. Think of all the websites you've been on, with snazzy-looking design, maybe with millions of dollars in sales passing through them, that want your password to be a mixture of uppercase and lowercase letters and numbers. In other words, they want you to enter "Password1!" instead of "correct horse battery staple". Every one of those websites is doing a thing that looks humorously silly to someone with a full security mindset or even just somebody who regularly reads XKCD. It says that the security system was set up by somebody who didn't know what they were doing and was blindly imitating impressive-looking mistakes they saw elsewhere.

Do you think that makes a good impression on their customers? That's right, it does! Because the customers don't know any better. Do you think that login system makes a good impression on the company's investors, including professional VCs and probably some angels with their own startup experience? That's right, it does! Because the VCs don't know any better, and even the angel doesn't know any better, and they don't realize they're missing a vital skill, and they aren't consulting anyone who knows more. An innocent is *impressed* if a website requires a mix of uppercase and lowercase letters and numbers *and* punctuation. They think the people running the website must really care to impose a security measure that unusual and inconvenient. The people running the website think that's what they're doing too.

People with deep security mindset are both rare and rarely *appreciated*. You can see just from the login system that none of the VCs and none of the C-level executives at

that startup thought they needed to consult a real professional, or managed to find a real professional rather than an empty suit if they went consulting. There was, visibly, nobody in the neighboring system with the combined knowledge and status to walk over to the CEO and say, “Your login system is embarrassing and you need to hire a real security professional.” Or if anybody did say that to the CEO, the CEO was offended and shot the messenger for not phrasing it ever-so-politely enough, or the CTO saw the outsider as a political threat and bad-mouthed them out of the game.

Your wishful should-universe hypothesis that people who can touch the full security mindset are more common than that within the venture capital and angel investing ecosystem is just flat wrong. Ordinary paranoia directed at widely-known adversarial cases is dense enough within the larger ecosystem to exert widespread social influence, albeit still comically absent in many individuals and regions. People with the full security mindset are too rare to have the same level of presence. That's the *easily visible* truth. You can *see* the login systems that want a punctuation mark in your password. You are not hallucinating them.

AMBER: If that's all true, then I just don't see how I can win. Maybe I should just condition on everything you say being false, since, if it's true, my winning seems unlikely—in which case all victories on my part would come in worlds with other background assumptions.

CORAL: … is that something you say often?

AMBER: Well, I say it whenever my victory starts to seem sufficiently unlikely.

CORAL: Goodness. I could maybe, *maybe* see somebody saying that once over the course of their entire lifetime, for a single unlikely conditional, but doing it more than once is sheer madness. I'd expect the unlikely conditionals to build up very fast and drop the probability of your mental world to effectively zero. It's tempting, but it's usually a bad idea to slip sideways into your own private hallucinatory universe when you feel you're under emotional pressure. I tend to believe that no matter what the difficulties, we are most likely to come up with good plans when we are mentally living in reality as opposed to somewhere else. If things seem difficult, we must face the difficulty squarely to succeed, to come up with some solution that faces down how bad the situation really is, rather than deciding to condition on things not being difficult because then it's too hard.

AMBER: Can you at least *try* talking to Mr. Topaz and advise him how to make things be secure?

CORAL: Sure. Trying things is easy, and I’m a character in a dialogue, so my opportunity costs are low. I'm sure Mr. Topaz is trying to build secure merchant drones, too. It's succeeding at things that is the hard part.

AMBER: Great, I'll see if I can get Mr. Topaz to talk to you. But do please be polite! If you think he's doing something wrong, try to point it out more gently than the way you've talked to me. I think I have enough political capital to get you in the door, but that won't last if you're rude.

CORAL: You know, back in mainstream computer security, when you propose a new way of securing a system, it's considered traditional and wise for everyone to gather around and try to come up with reasons why your idea might not work. It's understood that no matter how smart you are, most seemingly bright ideas turn out to be flawed, and that you shouldn't be touchy about people trying to shoot them down. Does Mr.

Topaz have no acquaintance at all with the practices in computer security? A lot of programmers do.

AMBER: I think he'd say he respects computer security as its own field, but he doesn't believe that building secure operating systems is the same problem as building merchant drones.

CORAL: And if I suggested that this case might be similar to the problem of building a secure operating system, and that this case creates a similar need for more effortful and cautious development, requiring both (a) additional development time and (b) a special need for caution supplied by people with unusual mindsets above and beyond ordinary paranoia, who have an unusual skill that identifies shaky assumptions in a safety story before an ordinary paranoid would judge a fire as being urgent enough to need putting out, who can remedy the problem using deeper solutions than an ordinary paranoid would generate as parries against imagined attacks?

If I suggested, indeed, that this scenario might hold generally wherever we demand robustness of a complex system that is being subjected to strong external or internal optimization pressures? Pressures that strongly promote the probabilities of particular states of affairs via optimization that searches across a large and complex state space? Pressures which therefore in turn subject other subparts of the system to selection for weird states and previously unenvisioned execution paths? Especially if some of these pressures may be in some sense creative and find states of the system or environment that surprise us or violate our surface generalizations?

AMBER: I think he'd probably think you were trying to look smart by using overly abstract language at him. Or he'd reply that he didn't see why this took any more caution than he was already using just by testing the drones to make sure they didn't crash or give out too much money.

CORAL: I see.

AMBER: So, shall we be off?

CORAL: Of course! No problem! I'll just go meet with Mr. Topaz and use verbal persuasion to turn him into Bruce Schneier.

AMBER: That's the spirit!

CORAL: God, how I wish I lived in the territory that corresponds to your map.

AMBER: Hey, come on. Is it seriously *that* hard to bestow exceptionally rare mental skills on people by talking at them? I agree it's a bad sign that Mr. Topaz shows no sign of wanting to acquire those skills, and doesn't think we have enough relative status to continue listening if we say something he doesn't want to hear. But that just means we have to phrase our advice cleverly so that he *will* want to hear it!

CORAL: I suppose you could modify your message into something Mr. Topaz doesn't find so unpleasant to hear. Something that sounds related to the topic of drone security, but which doesn't cost him much, and of course does not actually cause his drones to end up secure because that would be all unpleasant and expensive. You could slip a little sideways in reality, and convince yourself that you've gotten Mr. Topaz to ally with you, because he sounds agreeable now. Your instinctive desire for the high-status monkey to be on your political side will feel like its problem has been solved. You can substitute the feeling of having solved that problem for the unpleasant sense of not

having secured the actual drones; you can tell yourself that the bigger monkey will take care of everything now that he seems to be on your pleasantly-modified political side. And so you will be happy. Until the merchant drones hit the market, of course, but that unpleasant experience should be brief.

AMBER: Come on, we can do this! You've just got to think positively!

CORAL: … Well, if nothing else, this should be an interesting experience. I've never tried to do anything quite this doomed before.

Reason as Attentional Prosthesis

[*cross-posted from:* http://garybasin.com/reason-as-attentional-prosthesis/]

What are we doing when thinking, or *reasoning*?

Often, thinking is goal-directed. There is some outcome we are trying to achieve or problem we are trying to solve. In hindsight, answers often seem obvious. “Why didn’t *I* think of that?” For the most interesting problems, there is no formula or system that will lead to an optimal solution. The hard part becomes *framing* the problem. How do we do that? I have some thoughts on the role of attention.

But first, pigeons! In an experiment, pigeons are rewarded for pecking a toy banana hanging from the ceiling of their enclosure, but only while standing on a box placed below it. They’re also rewarded, in a separate setup with no toy banana present, for moving a box to randomly marked spots on the wall of their enclosure.

And then they’re faced with a novel puzzle:

> “When these pigeons were confronted with the banana out of reach, the box present but displaced from under it, and no spot was on the wall, **they looked back and forth between banana and box at first, appearing “confused” to human observers.** Then, within less than two minutes, they abruptlybegan to push the box in the direction of the banana, stopped when it was underneath, climbed and pecked.” (Epstein, Kirshnit, Lanza, & Rubin, 1984).

A conundrum that stresses the bird brain. Its attention is drawn to the banana, motivated by reward previously reinforced, and yet something is wrong. The pattern is not quite the same as before. Our feathered friend has yet to encounter a banana and a box at the same time where the box was not under the banana. Also, he has not been previously rewarded for moving a box except to a marked target location on the wall. **To solve this problem, the pigeon seems to need to somehow suppress its default, habituated behavior.** We can see it struggling to do so. Perhaps this is the pigeon-equivalent of “dropping preconceptions”, or our habitual **way of seeing** a situation. Once a primary reaction is suppressed, it can hold attention on other options which don’t immediately seem like answers to its problem.

Even more is demanded of the pigeon. He must figure out how to combine familiar behaviors in a new way. The pigeon experiences pleasure — anticipated reward — when he sees a box and a marked spot on the wall. Now, there is no marked spot. He also experiences pleasure when he sees a box under a banana. Now, the box is elsewhere. How to fit these pieces together to cause new behavior? Presumably, he is doing ***something like reasoning***. He is aware of the box, an ability (affordance) to move boxes, and the good feelings associated with a box under a banana. He must be able to ignore other aspects of the scene which are not relevant. **By first simplifying his experience to relevant aspects, he can attempt novel combinations.** Once perceiving a candidate solution, new action is a foregone conclusion — the pigeon is quick to push the box across the cage.

Reason seems to serve as a kind of ***attentional prosthesis***. By default, humans, and pigeons, are responding reflexively — according to habit. When reflexes aren’t good enough, we need some way to explore pieces of past experiences and recombine them in new ways. In the process, we need help navigating our attention away from

what is often the most attractive stimulus. Reason functions as such a mechanism. We recall memories of situations where only certain aspects (“abstractions” or “concepts”) are attended to. For each aspect, we trigger a re-cognition — a re-experiencing — of previous experiences that are *somehow* related. **The relation can be a similarity of stimulus, but also, crucially, similarity of outcome or response.** This is also known as *functional* or *mediated* generalization. By compressing into relevant aspects, the mind seems to be more capable of *chaining together* previously unrelated “sub-experiences”. If a chain is conceivable to an emotionally-appealing outcome, action can follow.

Open thread, November 21 - November 28, 2017

You can find the last Open Thread here.

If it's worth saying, but not worth its own post, then it goes here.

Status Regulation and Anxious Underconfidence

Follow-up to: Against Modest Epistemology

I've now given my critique of modesty as a set of explicit doctrines. I've tried to give the background theory, which I believe is nothing more than conventional cynical economics, that explains why so many aspects of the world are not optimized to the limits of human intelligence in the manner of financial prices. I have argued that the essence of rationality is to adapt to whatever world you find yourself in, rather than to be "humble" or "arrogant" *a priori*. I've tried to give some preliminary examples of how we *really, really* don't live in the Adequate World where constant self-questioning would be appropriate, the way it *is* appropriate when second-guessing equity prices. I've tried to systematize modest epistemology into a semiformal rule, and I've argued that the rule yields absurd consequences.

I was careful to say all this first, because there's a strict order to debate. If you're going to argue against an idea, it's bad form to start off by arguing that the idea was generated by a flawed thought process, before you've explained why you think the idea itself is wrong. Even if we're refuting geocentrism, we should first say how we know that the Sun does not orbit the Earth, and *only then* pontificate about what cognitive biases might have afflicted geocentrists. As a rule, an idea should initially be discussed as though it had descended from the heavens on a USB stick spontaneously generated by an evaporating black hole, before any word is said psychoanalyzing the people who believe it. Otherwise I'd be guilty of poisoning the well, also known as Bulverism.

But I've now said quite a few words about modest epistemology as a pure idea. I feel comfortable at this stage saying that I think modest epistemology's popularity owes something to its emotional appeal, as opposed to being strictly derived from epistemic considerations. In particular: emotions related to social status and self-doubt.

Even if I thought modesty were the correct normative epistemology, I would caution people not to confuse the correct reasoning principle with those particular emotional impulses. You'll observe that I've written one or two things above about how *not* to analyze inadequacy, and mistakes not to make. We hear far too little from its advocates about potential misuses and distortions of modest epistemology, if we're going to take modest epistemology seriously as a basic reasoning mode, technique, or principle.

And I'll now try to describe the kinds of feelings that I think modesty's appeal rests on. Because I've come to appreciate increasingly that human beings are *really genuinely* different from one another, you shouldn't be surprised if it seems to you like this is *not* how you work. I claim nonetheless that many people do work like this.

i.

Let's start with the emotion—not restricted to cases of modesty, just what I suspect to be a common human emotion—of "anxious underconfidence."

As I started my current writing session, I had just ten minutes ago returned from the following conversation with someone looking for a job in the Bay Area that would give them relevant experience for running their own startup later:

ELIEZER: Are you a programmer?

ASPIRING FOUNDER: That's what everyone asks. I've programmed at all of my previous jobs, but I wouldn't call myself a programmer.

ELIEZER: I think you should try asking (person) if they know of any startups that could use non-super programmers, and look for a non-doomed startup that's still early-stage enough that you can be assigned some business jobs and get a chance to try your hand at that without needing to manage it yourself. That might get you the startup experience you want.

ASPIRING FOUNDER: I know how to program, but I don't know if I can display that well enough. I don't have a Github account. I think I'd have to spend three months boning up on programming problems before I could do anything like the Google interview—or maybe I could do one of the bootcamps for programmers—

ELIEZER: I'm not sure if they're aimed at your current skill level. Why don't you try just one interview and see how that goes before you make any complicated further plans about how to prove your skills?

This fits into a very common pattern of advice I've found myself giving, along the lines of, "Don't assume you can't do something when it's very cheap to try testing your ability to do it," or, "Don't assume other people will evaluate you lowly when it's cheap to test that belief."

I try to be careful to distinguish the virtue of avoiding overconfidence, which I sometimes call "humility," from the phenomenon I'm calling "modest epistemology." But even so, when overconfidence is such a terrible scourge according to the cognitive bias literature, can it ever be wise to caution people against *under*confidence?

Yes. First of all, overcompensation after being warned about a cognitive bias is also a recognized problem in the literature; and the literature on that talks about how bad people often are at determining whether they're undercorrecting or overcorrecting.[1] Second, my own experience has been that while, yes, commenters on the Internet are often overconfident, it's very different when I'm talking to people in person. My more recent experience seems more like 90% telling people to be less underconfident, to reach higher, to be more ambitious, to test themselves, and maybe 10% cautioning people against overconfidence. And yes, this ratio applies to men as well as women and nonbinary people, and to people considered high-status as well as people considered low-status.

Several people have now told me that the most important thing I have ever said to them is: “If you *never* fail, you’re only trying things that are too easy and playing far below your level.” Or, phrased as a standard Umeshism: “If you can’t remember any time in the last six months when you failed, you aren’t trying to do difficult enough things.” I first said it to someone who had set themselves on a career track to becoming a nurse instead of a physicist, even though they liked physics, because they were *sure* they could succeed at becoming a nurse.

I call this “anxious underconfidence,” and it seems to me to share a common thread with social anxiety. We might define “social anxiety” as “experiencing fear far in excess of what a third party would say are the reasonably predictable exterior consequences, with respect to other people possibly thinking poorly of you, or wanting things from you that you can’t provide them.” If someone is terrified of being present at a large social event because someone there might *talk* to them and they might be confused and stutter out an answer—when, realistically, this at worst makes a transient poor impression that is soon forgotten because you are not at the center of the other person’s life—then this is an *excess* fear of that event.

Similarly, many people’s emotional makeup is such that they experience what I would consider an excess fear—a fear disproportionate to the non-emotional consequences—of *trying something and failing*. A fear so strong that you become a nurse instead of a physicist because that is something you are *certain* you can do. Anything you might *not* be able to do is crossed off the list instantly. In fact, it was probably never generated as a policy option in the first place. Even when the correct course is obviously to just try the job interview and see what happens, the test will be put off indefinitely if failure feels possible.

If you’ve never wasted an effort, you’re filtering on far too high a required probability of success. Trying to avoid wasting effort—yes, that’s a good idea. Feeling bad when you realize you’ve wasted effort—yes, I do that too. But some people slice off the entire realm of uncertain projects because the prospect of having wasted effort, of having been publicly wrong, seems so horrible that projects in this class are not to be considered.

This is one of the emotions that I think might be at work in recommendations to take an outside view on your chances of success in some endeavor. If you only try the things that are allowed for your “reference class,” you’re supposed to be safe—in a certain social sense. You may fail, but you can justify the attempt to others by noting that many others have succeeded on similar tasks. On the other hand, if you try something more ambitious, *you could fail and have everyone think you were stupid to try*.

The mark of this vulnerability, and the proof that it is indeed a fallacy, would be not *testing* the predictions that the modest point of view makes about your inevitable failures—even when they would be cheap to test, and even when failure doesn’t lead to anything that a non-phobic third party would rate as terrible.

ii.

The other emotions I have in mind are perhaps easiest to understand in the context of efficient markets.

In humanity's environment of evolutionary adaptedness, an offer of fifty carrots for a roasted antelope leg reflects a judgment about roles, relationships, and status. This idea of "price" is easier to grasp than the economist's notion; and given that somebody *doesn't* have the economist's very specific notion in mind when you speak of "efficient markets," they can end up making what I would consider an extremely understandable mistake.

You tried to explain to them that even if they thought AAPL stock was underpriced, they ought to question themselves. You claimed that they *couldn't* manage to be systematically right on the occasions where the market price swung drastically. Not unless they had access to insider information on single stocks—which is to say, they just couldn't do it.

But "I can't do that. *And you can't either!*" is a suspicious statement in everyday life. Suppose I try to juggle two balls and succeed, and then I try to juggle three balls and drop them. I could conclude that I'm bad at juggling and that other people could do better than me, which comes with a loss of status. Alternatively, I could heave a sad sigh as I come to realize that juggling more than two balls is just not possible. Whereupon my social standing in comparison to others is preserved. I even get to give instruction to others about this hard-won life lesson, and smile with sage superiority at any young fools who are still trying to figure out how to juggle three balls at a time.

I grew up with this fallacy, in the form of my Orthodox Jewish parents smiling at me and explaining how when they were young, they had asked a lot of religious questions too; but then they grew out of it, coming to recognize that some things were just beyond our ken.

At the time, I was flabbergasted at my parents' arrogance in assuming that because they couldn't solve a problem as teenagers, nobody else could possibly solve it going forward. Today, I understand this viewpoint not as arrogance, but as a simple flinch away from a painful thought and toward a pleasurable one. You can admit that you failed where success was possible, or you can smile with gently forgiving superiority at the youthful enthusiasm of those who are still naive enough to attempt to do better.

Of course, some things *are* impossible. But if one's flinch response to failure is to perform a mental search for reasons one couldn't have succeeded, it can be tempting to slide into false despair.

In the book *Superforecasting*, Philip Tetlock describes the number one characteristic of top forecasters, who show the ability to persistently outperform professional analysts and even small prediction markets: *they believe that outperformance in forecasting is possible, and work to improve their performance*.[2]

I would expect this to come as a shock to people who grew up steeped in academic studies of overconfidence and took away the lesson that epistemic excellence is mostly about accepting your own limitations.[3] But I read that chapter of *Superforecasting* and laughed, because I was pretty sure from my own experience that I could guess what had happed to Tetlock: he had run into large numbers of naive respondents who smiled condescendingly at the naive enthusiasm of those who thought that anyone can get good at predicting future events.[4]

Now, imagine you're somebody who didn't read *Superforecasting*, but did at least grow up with parents telling you that if they're not smart enough to be a lawyer, then

neither are you. (As happened to a certain childhood friend of mine who is now a lawyer.)

And then you run across somebody who tries to tell you, not just that *they* can't outguess the stock market, but that *you're* not allowed to become good at it either. They claim that nobody is allowed to master the task at which they failed. Your uncle tripled his savings when he bet it all on GOOG, and this person tries to wave it off as luck. Isn't that like somebody condescendingly explaining why juggling three balls is impossible, after you've seen with your own eyes that your uncle can juggle four?

This isn't a naive question. Somebody who has seen the condescension of despair in action is right to treat this kind of claim as suspicious. It *ought* to take a massive economics literature examining the idea in theory and in practice, and responding to various apparent counterexamples, before we accept that a new kind of near-impossibility has been established in a case where the laws of physics seem to leave the possibility open.

Perhaps what you said to the efficiency skeptic was something like:

> If it's obvious that AAPL stock should be worth more because iPhones are so great, then a hedge fund manager should be able to see this logic too. This means that this information will already be baked into the market price. If what you're saying is true, the market already knows it—and what the market knows beyond that, neither you nor I can guess.

But what they *heard* you saying was:

> O thou, who burns with tears for those who burn,
> In Hell, whose fires will find thee in thy turn
> Hope not the Lord thy God to mercy teach
> For who art thou to teach, or He to learn?[5]

This again is an obvious fallacy for them to suspect you of committing. They're suggesting that something might be wrong with *Y*'s judgment of *X*, and you're telling them to shut up because *Y* knows far better than them. Even though you can't point to any flaws in the skeptic's suggestion, can't say anything about the kinds of reasons *Y* has in mind for believing *X*, and can't point them to the information sources *Y* might be drawing from. And it just so happens that *Y* is big and powerful and impressive.

If we could look back at the ages before liquid financial markets existed, and record all of the human conversations that went on at the time, then practically every instance in history of anything that sounded like what you said about efficient markets—that some mysterious powerful being is always unquestionably right, though the reason be impossible to understand—would have been a mistake or a lie. So it's hard to blame the skeptic for being suspicious, if they don't yet understand how market efficiency works.

What you said to the skeptic about AAPL stock is justified for extremely liquid markets on short-term time horizons, but—at least I would claim—very rarely justified anywhere else. The claim is, "If you think you know the price of AAPL better than the stock market, then no matter how good the evidence you think you've found is, your reasoning just has some hidden mistake, or is neglecting some unspecified key consideration." And no matter how valiantly they argue, no matter how carefully they construct their reasoning, we just smile and say, "Sorry, kid." It is a final and absolute

slapdown that is *meant* to be inescapable by any mundane means within a common person's grasp.

Indeed, this supposedly inescapable and crushing rejoinder looks surprisingly similar to a particular social phenomenon I'll call "status regulation."

iii.

Status is an extremely valuable resource, and was valuable in the ancestral environment.

Status is also a somewhat conserved quantity. Not everyone can be sole dictator.

Even if a hunter-gatherer tribe or a startup contains more average status per person than a medieval society full of downtrodden peasants, there's still a sense in which status is a limited resource and letting someone walk off with lots of status is like letting them walk off with your bag of carrots. So it shouldn't be surprising if *acting like you have more status than I assign to you* triggers a negative emotion, a slapdown response.

If slapdowns exist to limit access to an important scarce resource, we should expect them to be cheater-resistant in the face of intense competition for that resource.[6] If just anyone could find some easy sentences to say that let them get higher status than God, then your system for allocating status would be too easy to game. Escaping slapdowns should be hard, generally requiring more than mere abstract argumentation.

Except that *people are different*. So not everyone feels the same way about this, any more than we all feel the same way about sex.

As I've increasingly noticed of late, and contrary to beliefs earlier in my career about the psychological unity of humankind, not all human beings have all the human emotions. The logic of sexual reproduction makes it unlikely that anyone will have a new complex piece of mental machinery that nobody else has… but *absences* of complex machinery aren't just possible; they're amazingly common.

And we tend to underestimate how different other people are from ourselves. Once upon a time, there used to be a great and acrimonious debate in philosophy about whether people had "mental imagery" (whether or not people actually see a little picture of an elephant when they think about an elephant). It later turned out that some people see a little picture of an elephant, some people don't, and both sides thought that the way they personally worked was so fundamental to cognition that they couldn't imagine that other people worked differently. So both sides of the philosophical debate thought the other side was just full of crazy philosophers who were willfully denying the obvious. The typical mind fallacy is the bias whereby we assume most other people are much more like us than they actually are.

If you're fully asexual, then you haven't felt the emotion others call "sexual desire"… but you can feel friendship, the warmth of cuddling, and in most cases you can experience orgasm. If you're not around people who talk explicitly about the possibility of asexuality, you might not even realize you're asexual and that there is a

distinct “sexual attraction” emotion you are missing, just like some people with congenital anosmia never realize that they don’t have a sense of smell.

Many people seem to be the equivalent of asexual with respect to the emotion of status regulation—myself among them. If you’re blind to status regulation (or even status itself) then you might still see that people with status get respect, and hunger for that respect. You might see someone with a nice car and envy the car. You might see a horrible person with a big house and think that their behavior ought not to be rewarded with a big house, and feel bitter about the smaller house you earned by being good. I can feel all of those things, but people’s overall place in the pecking order isn’t a fast, perceptual, pre-deliberative *thing* for me in its own right.

For many people, I gather that the social order is a reified emotional *thing* separate from respect, separate from the goods that status can obtain, separate from any deliberative reasoning about who ought to have those goods, and separate from any belief about who consented to be part of an implicit community agreement. There’s just a felt sense that some people are lower in various status hierarchies, while others are higher; and overreaching by trying to claim significantly more status than you currently have is an offense against the reified social order, which has an immediate emotional impact, separate from any beliefs about the further consequences that a social order causes. One may *also* have explicit beliefs about possible benefits or harms that could be caused by disruptions to the status hierarchy, but the status regulation feeling is more basic than that and doesn’t depend on high-level theories or cost-benefit calculations.

Consider, in this context, the efficiency skeptic’s perspective:

SKEPTIC: I have to say, I'm baffled at your insistence that hedge fund managers are the summit of worldly wisdom. Many hedge fund managers—possibly most—are nothing but charlatans who convince pension managers to invest money that ought to have gone into index funds.

CECIE: Markets are a mechanism that allow and incentivize a single smart participant to spot a bit of free energy and eat it, in a way that aggregates to produce a global equilibrium with no free energy. We don’t need to suppose that most hedge fund managers are wise; we only need to suppose that a tiny handful of market actors are smart enough in each case to have already seen what you saw.

SKEPTIC: I’m not sure I understand. It sounds like what you’re saying, though, is that your faith is not in mere humans, but in some mysterious higher force, the “Market.”

You consider this Market incredibly impressive and powerful. You consider it folly for anyone to think that they can know better than the Market. And you just happen to have on hand a fully general method for slapping down anyone who dares challenge the Market, without needing to actually defend this or that particular belief of the Market.

CECIE: A market’s efficiency doesn’t derive from its social status. True efficiency is very rare in human experience. There’s a very good reason that we had to coin a term for the concept of “efficient markets,” and not “efficient medicine” or “efficient physics”: because in those ecologies, not just anyone can come along and consume a morsel of free energy.

If you personally know better than the doctors in a hospital, you can't walk in off the street tomorrow and make millions of dollars saving more patients' lives. If you personally know better than an academic field, you can't walk in off the street tomorrow and make millions of dollars filling the arXiv with more accurate papers.

SKEPTIC: I don't know. The parallels between efficiency and human status relations seem awfully strong, and this "Market moves in mysterious ways" rejoinder seems like an awfully convenient trick.

Indeed, I would be surprised if there *weren't* at least some believers in "efficient markets" who assigned them extremely high status and were tempted to exaggerate their efficiency, perhaps feeling a sense of indignation at those who dared to do better. Perhaps there are people who feel an urge to slap down anyone who starts questioning the efficiency of Boomville's residential housing market.

So be it; Deepak Chopra can't falsify quantum mechanics by being enthusiastic about a distorted version of it. The efficiency skeptic should jettison their skepticism, and should take care to avoid the fallacy fallacy—the fallacy of taking for granted that some conclusion is false just because a fallacious argument for that conclusion exists.[7]

I once summarized my epistemology like so: "Try to make sure you'd arrive at different beliefs in different worlds." You don't want to think in such a way that you wouldn't believe in a conclusion in a world where it were true, just because a fallacious argument could support it. Emotionally appealing mistakes are not invincible cognitive traps that nobody can ever escape from. Sometimes they're not even that hard to escape.

The remedy, as usual, is technical understanding. If you know in detail when a phenomenon switches on and off, and when the "inescapable" slapdown is escapable, you probably won't map it onto God.

iv.

I actually can't recall seeing anyone make the mistake of treating efficient markets like high-status authorities in a social pecking order.[8] The more general phenomenon seems quite common, though: heavily weighting relative status in determining odds of success; responding to overly ambitious plans as though they were not merely imprudent but impudent; and privileging the hypothesis that authoritative individuals and institutions have mysterious unspecified good reasons for their actions, even when these reasons stubbornly resist elicitation and the actions are sufficiently explained by misaligned incentives.

From what I can tell, status regulation is a second factor accounting for modesty's appeal, distinct from anxious underconfidence. The impulse is to construct "cheater-resistant" slapdowns that can (for example) prevent dilettantes who are low on the relevant status hierarchy from proposing new SAD treatments. Because if dilettantes can exploit an inefficiency in a respected scientific field, then this makes it easier to "steal" status and upset the current order.

In the past, I didn't understand that an important part of status regulation, as most people experience it, is that one needs to already possess a certain amount of status before it's seen as acceptable to reach up for a given higher level of status. What could be wrong (I previously thought) with trying to bestow unusually large benefits upon your tribe? I could understand why it would be bad to claim that you had already accomplished more than you had—to claim more respect than was due the good you'd already done. But what could be wrong with trying to do more good for the tribe, in the future, than you already had in the present?

It took me a long time to understand that *trying to do interesting things in the future* is a status violation because your current status right now determines what kinds of images you are allowed to associate with yourself, and if your status is low, then many people will intuitively perceive an unpleasant violation of the social order should you associate with yourself an image of possible future success above some level. Only people who already have something like an aura of *pre*-importance are allowed to try to do important things. Publicly setting out to do valuable and important things *eventually* is above the status you already have *now*, and will generate an immediate system-1 slapdown reaction.

I recognize now that this is a common lens through which people see the world, though I still don't know how it feels to *feel* that.

Regardless, when I see a supposed piece of epistemology that looks to me an *awful* lot like my model of status regulation, but which doesn't seem to cohere with the patterns of correct reasoning described by theorists like E. T. Jaynes, I get suspicious. When people cite the "outside view" to argue that one should stick to projects whose ambition and impressiveness befit one's "reference class," and announce that any effort to significantly outperform the "reference class" is epistemically suspect "overconfidence," and insist that moving to take into account local extenuating factors, causal accounts, and justifications constitutes an illicit appeal to the "inside view" and we should rely on more obvious, visible, *publicly demonstrable* signs of *overall auspiciousness or inauspiciousness*… you know, I'm not sure this is strictly inspired by the experimental work done on people estimating their Christmas shopping completion times.

I become suspicious as well when this model is deployed in practice by people who talk in the same tone of voice that I've come to associate with status regulation, and when an awful lot of what they say sounds to me like an elaborate rationalization of, "Who are *you* to act like some kind of big shot?"

I observe that many of the same people worry a lot about "What do you say to the Republican?" or the possibility that crackpots might try to *cheat*—like they're trying above all to guard some valuable social resource from the possibility of theft. I observe that the notion of somebody being able to steal that resource *and get away with it* seems to inspire a special degree of horror, rather than just being one more case of somebody making a mistaken probability estimate.

I observe that *attempts to do much better than is the norm* elicit many heated accusations of overconfidence. I observe that *failures to even try to live up to your track record or to do as well as a typical member of some suggested reference class* mysteriously fail to elicit many heated accusations of *under*confidence. Underconfidence and overconfidence are symmetrical mistakes *epistemically*, and yet somehow I never see generalizations of the outside view even-handedly applied to correct both biases.

And so I’m skeptical that this reflects normative probability theory, pure epistemic rules such as aliens would also invent and use. Sort of like how an asexual decision theorist might be skeptical of an argument saying that the pure structure of decision theory implies that arbitrary decision agents with arbitrary biologies ought to value sex.

This kind of modesty often looks like the condescension of despair, or bears the “God works in mysterious ways” property of attributing vague good reasons to authorities on vague grounds. It’s the kind of reasoning that makes sense in the context of an efficient market, but it doesn’t seem to be coming from a model of the structure or incentives of relevant communities, such as the research community studying mood disorders.

No-free-energy equilibria do generalize beyond asset prices; markets are not the only ecologies full of motivated agents. But sometimes those agents aren’t sufficiently motivated and incentivized to do certain things, or the agents aren’t all individually free to do them. In this case, I think that many people are doing the equivalent of humbly accepting that they can’t possibly know whether a single house in Boomville is overpriced. In fact, I think this form of status-oriented modesty is extremely common, and is having hugely detrimental effects on the epistemic standards and the basic emotional health of the people who fall into it.

v.

Modesty can take the form of an explicit epistemological norm, or it can manifest in more quiet and implicit ways, as small flinches away from painful thoughts and towards more comfortable ones. It’s the latter that I think is causing most of the problem. I’ve spent a significant amount of time critiquing the explicit norms, because I think these serve an important role as canaries piling up in the coalmine, and because they are bad epistemology in their own right. But my chief hope is to illuminate that smaller and more quiet problem.

I think that anxious underconfidence and status regulation are the main forces motivating modesty, while concerns about overconfidence, disagreement, and theoreticism serve a secondary role in justifying and propagating these patterns of thought. Nor are anxious underconfidence and status regulation entirely separate problems; bucking the status quo is particularly painful when public failure is a possibility, and shooting low can be particularly attractive when it protects against accusations of hubris.

Consider the outside view as a heuristic for minimizing the risk of social transgression and failure. Relying on an outside view instead of an inside view will generally mean making fewer knowledge claims, and the knowledge claims will generally rest on surface impressions (which are easier to share), rather than on privileged insights and background knowledge (which imply more status).

Or consider the social utility of playing the fox's part. The fox can say that they rely only on humble data sets, disclaiming the hedgehog’s lofty theories, and disclaiming any special knowledge or special powers of discernment implied thereby. And by sticking to relatively local claims, or only endorsing global theories once they command authorities’ universal assent, the fox can avoid endorsing the kinds of

generalizations that might encroach on someone else's turf or otherwise disrupt a status hierarchy.

Finally, consider appeals to agreement. As a matter of probability theory, perfect rationality plus mutual understanding often entails perfect agreement. Yet it doesn't follow from this that the way for human beings to become more rational is to try their best to minimize disagreement. An all-knowing agent will assign probabilities approaching 0 and 1 to all or most of its beliefs, but this doesn't imply that the best way to become more knowledgeable is to manually adjust one's beliefs to be as extreme as possible.

The behavior of ideal Bayesian reasoners is important evidence about how to become more rational. What this usually involves, however, is understanding how Bayesian reasoning works internally and trying to implement a causally similar procedure, not looking at the end product and trying to pantomime particular surface-level indicators or side-effects of good Bayesian inference. And a psychological drive toward automatic deference or self-skepticism isn't the *mechanism* by which Bayesians end up agreeing to agree.

Bayes-optimal reasoners don't Aumann-agree because they're following some exotic meta-level heuristic. I don't know of any general-purpose rule like that for quickly and cheaply leapfrogging to consensus, except ones that do so by sacrificing some amount of expected belief accuracy. To the best of my knowledge, the outlandish and ingenious trick that really lets flawed reasoners inch nearer to Aumann's ideal is just the old-fashioned one where you go out and think about yourself and about the world, and do what you can to correct for this or that bias in a case-by-case fashion.

Whether applied selectively or consistently, the temptation of modesty is to "fake" Aumann agreement—to rush the process, rather than waiting until you and others can *actually* rationally converge upon the same views. The temptation is to call an early halt to risky lines of inquiry, to not claim to know too much, and to not claim to aspire to too much; all while wielding a fully general argument against anyone who doesn't do the same.

And now that I've given my warning about these risks and wrong turns, I hope to return to other matters.

My friend John thought that there were hidden good reasons behind Japan's decision not to print money. Was this because he thought that the Bank of Japan was big and powerful, and therefore higher status than a non-professional-economist like me?

I literally had a bad taste in my mouth as I wrote that paragraph.[9] This kind of psychologizing is not what people epistemically virtuous enough to bet on their beliefs should spend most of their time saying to one another. They should just be winning hundreds of dollars off of me by betting on whether some AI benchmark will be met by a certain time, as my friend later proceeded to do. And then later he and I both lost money to other friends, betting against Trump's election victory. The journey goes on.

I'm not scheming to taint all humility forever with the mere suspicion of secretly fallacious reasoning. That would convict me of the fallacy fallacy. Yes, subconscious influences and emotional temptations are a problem, but you can often beat those if your explicit verbal reasoning is good.

I've critiqued the fruits of modesty, and noted my concerns about the tree on which they grow. I've said why, though my understanding of the mental motions behind modesty is very imperfect and incomplete, I do not expect these motions to yield good and true fruits. But cognitive fallacies are not invincible traps; and if I spent most of my time thinking about meta-rationality and cognitive bias, I'd be taking my eye off the ball.[10]

Inadequate Equilibria is now available in electronic and print form on equilibriabook.com.

Conclusion: **Against Shooting Yourself in the Foot**.

---

1. From Bodenhausen, Macrae, and Hugenberg (2003):

   > [I]f correctional mechanisms are to result in a less biased judgment, the perceiver must have a generally accurate lay theory about the direction and extent of the bias. Otherwise, corrections could go in the wrong direction, they could go insufficiently in the right direction, or they could go too far in the right direction, leading to overcorrection. Indeed, many examples of overcorrection have been documented (see Wegener & Perry, 1997, for a review), indicating that even when a bias is detected and capacity and motivation are present, controlled processes are not necessarily effective in accurately counteracting automatic biases. ↩

2. From *Superforecasting*: "The strongest predictor of rising into the ranks of superforecasters is perpetual beta, the degree to which one is committed to belief updating and self-improvement. It is roughly three times as powerful a predictor as its closest rival, intelligence." ↩

3. E.g., Alpert and Raiffa (1982), "A Progress Report on the Training of Probability Assessors. ↩

4. Or rather, get better at predicting future events than intelligence agencies, company executives, and the wisdom of crowds. ↩

5. From Edward FitzGerald's *Rubaiyat of Omar Khayyám*. ↩

6. The existence of specialized cognitive modules for detecting cheating can be seen, e.g., in the Wason selection task. Test subjects perform poorly when asked to perform a version of this task introduced in socially neutral terms (e.g., rules governing numbers and colors), but perform well when given an isomorphic version of the task that is framed in terms of social rules and methods for spotting violators of those rules. See Cosmides and Tooby, "Cognitive Adaptations for Social Exchange." ↩

7. Give me any other major and widely discussed belief from any other field of science, and I shall paint a picture of how it resembles some other fallacy—maybe even find somebody who actually misinterpreted it that way. It doesn't mean much. There's just such a vast array of mistakes human minds can make that if you rejected every argument that looks like it could maybe be guilty of some fallacy, you'd be left with nothing at all.

   It often just doesn't mean very much when we find that a line of argument can be made to look "suspiciously like" some fallacious argument. Or rather: being suspicious is one thing, and being

so suspicious that relevant evidence cannot realistically overcome a suspicion is another. ↩

8. It's a mistake that somebody could make, though, and people promoting ideas that are susceptible to fallacious misinterpretation do have an obligation to post warning signs. Sometimes it feels like I've spent my whole life doing nothing else. ↩

9. Well, my breakfast might also have had something to do with it, but I *noticed* the bad taste while writing those sentences. ↩

10. There's more I can say about how I think modest epistemology and status dynamics work in practice, based on past conversations; but it would require me to digress into talking about my work and fiction-writing. For a supplemental chapter taking a more concrete look at these concepts, see Hero Licensing. ↩

Normative assumptions: answers, emotions, and narratives

Agents can be modelled as pairs (p, R), where R is a reward, and p is a planner that takes R and more or less rationally outputs the policy p(R).

Normative assumptions are assumptions that can distinguish between different pairs (p, R) and (p', R'), despite those pairs generating the same policy p(R) = p'(R'), and hence being indistinguishable by observation. We've already looked at the normative assumption "stated regret is accurate"; this post will look at more general normative assumptions, and why you'd want to use them to define human (ir)rationality and reward.

Talking it all out

The previous post used "stated reward". This suggests an extension: why restrict to certain human utterances like regret, and why not just use human utterances in general? If we want to know about human values and human rationality, why not ask... humans?

But humans, and I'm sorry if this shocks you, sometimes lie or mislead or tell less than the full truth. We're subject to biases and inconsistencies, and hence our opinions are not a reliable indicator of our values.

Can we not just detect biases and lies, and train a learning agent to ignore or discount them? For example, we could use labelled databases of lies or partial truths, and...

The problem is that a "labelled database of lies and partial truths" is just asking humans again, at a meta level. If we can be misleading at the object level, how can we trust the meta-level assessment ("I am not racist"; "I am telling the truth in the previous statement")? If we use different people to do the speaking and the labelling, then we're just using the biases and utterance of some people as assessments of the biases and utterances of others.

Especially when people's values are so under-defined, and biases are both rampant and un-obvious, and we expect meta-level statements to be even more noisy that object level ones.

Let P be the procedure ="take a lot of human statements about values, selected according to some criteria, and a lot of human meta-statements about values and object level statements, do some sort of machine learning on them with some sort of regularisation". In the language of this post, is P something that you would be comfortable to see as the *definition* of human values?

As definitions go, it certainly seems like it's not completely useless, but it probably has some hideous edge case failures. Can we try and do better?

Why is bias bias, and rationality rationality?

Returning to the point of previous posts, one cannot deduce human reward and rationality from observations. Humans, however, do it all the time, about themselves and about each other, and we often agree with each other. So how do we do it?

Basically, we need to add normative assumptions - assumptions, not derived from observations about the world, that distinguish between different models of (ir)rationality and reward.

That old post mentioned feelings of regret ("I shouldn't have done that!"), which seem to be one of the prime reasons we model ourselves as having certain rewards - generally, the opposite to whatever we're regretting. When we specifically start regretting our own actions, rather than the outcome ("I knew it was the wrong decision when I made it!" "Why didn't I stop to think?") this helps us model ourselves as partially irrational agents.

What other assumptions can we reach for? Basically, we need to look for how humans define rationality and irrationality, and use this to define our very values.

Rationality: logic and irrelevant elements?

One strong assumption that underlies the definition of rationality, is that it follows the rules of logic. By transitivity, if I prefer A to B, and B to C, then I must prefer A to C. In fact, an even more basic assumption is that people actually prefer A to B, or vice versa, or rank them equally.

So far, so good. But nobody is shocked that I prefer cereals to curry most mornings, and the reverse at all lunches. Do I not have a preference? Ah, but my preferences is time and appetite dependent, and people don't see this as a failure of rationality.

What we see as a failure of rationality, is if someone shifts behaviours for unimportant or irrelevant reasons. We put someone in a situation, check their A vs B preferences, put the same or a different person in a *functionally identical* situation, check their B vs C, etc...

Thus a lot of rationality can be reduced to logic, plus a theory about what constitutes a functionally identical situation. In other words, rationality is mainly a theory about what *doesn't matter*, or about what *shouldn't matter*.

Anchoring bias, emotions, and narratives

Let's look again at anchoring bias. In this bias, people who hear a low-but-irrelevant number will offer less for a product than people who hear a high-but-irrelevant number.

It's one of the biases that people agree the most is a bias - I've yet to hear anyone argue that people value possessing products at prices close to random numbers they're heard recently.

Now imagine the situation in which these tests would occur. It might be in a classroom, or a university room. Suggestive words might be "calm, dispassionate, clinical, formal, uniform" or terms of that nature.

Let's try another variant of the anchoring bias. In this version, the vendor either insults the subject, or compliments them. I'd be willing to bet that we would find people willing to pay more in the second case than in the first.

Finally, in the third variant, the participants are told that the whole interaction, including the insults/compliments and any subsequent decision, will be broadcast to the world.

Now, we have the same behaviour in three situations - random number, differential treatment, and public differential treatment. I'd classify the first as a bias, and the last as perfectly rational behaviour, with the middle situation falling somewhat in between.

What this means is that we judge that random numbers in relaxed environments are irrelevant to human values; that emotional interactions are relevant; and that publically visible emotional reactions are very relevant.

At this point I'd introduce narratives - the stories we tell ourselves, about ourselves and others. A strong element of these narratives is that we have few complex emotions and preferences, rather than many simple ones. Think of all the situations that go under the label "shame", "joy", or "doubt", for instances (and when it was more popular, all the factors that defined "honour").

Small isolated preferences get classified as quirks, or compulsions, and we generally feel that we could easily get rid of these without affecting the core of our personalities.

Back to the anchoring biases. In the original setting, everything is carefully constructed to remove strong emotions and preferences. We are not getting insulted. The setting is relaxed. The random numbers are not connected with anything deep about us. Therefore, according to our narratives, we have no genuine preferences or emotions at stake (apart from the preference for whatever is being sold). So the different behaviours that happen cannot be due to preferences, and must be biases.

In the other settings, we have strong emotions and then social judgement at stake, and our narratives about ourselves count these as very valid sources of preferences.

It's all a bit vague...

This mention of regret, emotions, and narratives seems a bit vague and informal. I haven't even bothered to provide any references, and there are weasel words like "generally". Could someone else not give different interpretations about what's going on, maybe by triggering invoking narratives about ourselves?

And indeed they could. The fundamental problem remains: our values are inconsistent and underdefined, and our narratives are also inconsistent and underdefined.

We still have to make choices about how to resolve these inconsistencies and definitions, and different people and different cultures would resolve them very differently.

Nevertheless, I think we have made some progress.

It's *all* humans answering questions - which ones?

If we make use of emotions and narratives, this suggests a slightly different way of proceeding. Rather than just taking a lot of human answers and meta-answers, we first identify the main emotions and narratives that we care about. We train the AIs to recognise these in humans (it need not be perfect - humans are not perfect at it either, but are better than chance).

Only then do we unleash it on human answers, and allow it to ask questions itself. Instead of drawing categories from the answers, we draw some of the categories ourselves first, and it then uses there to interpret the answers. We don't have to label which answers and meta-answers are true or reliable - the AI will draw its own inferences about that.

Now, you might argue that this is just human answering questions all over again - identifying human emotions is just the AI learning from labelled data, and the label is a human answer.

To which I'd say... of course it is. Everything we train an AI on is essentially human answers of some sort. We provide the training data and the labels, we tweak the parameters depending on results, we judge which method performs better.

But of course, in machine learning, some approaches are better than others, even if they're "equivalent" in this way. Human feedback is more useful distinguishing between some categories than between others. And it seems to me that "identifying strong narratives and emotions, training the AI on it, then unleashing it on a lot of examples of human (meta-)answers, and getting human feedback on its next level conclusions (often using the initial categories to phrase questions)" is a better and more stable approach than "unleash the AI on a lot of examples of human (meta-)answers, and let it draw its own categories".

For a start, I suspect the first approach will give a regularisation we're more comfortable with (as our narrative categories give some hints as to what we consider important and what we consider contingent). Secondly, I feel that this approach is less likely to collapse into pathological behaviour. And I feel that this could succeed at the task of "identify subconscious aspects of human preferences that we really want to endorse, if we but knew about them".

Productivity: Working towards a summary of what we know

Epistemic effort

- Spent 20-30 hours researching what we know about productivity. About an hour a day for two weeks, plus a few more days where I dedicated more than an hour to studying productivity.
- Skimmed through the entire archives of Cal Newport's blog (the author of Deep Work, which I had read about a year or two ago), and read the posts that seem most useful.
- Searched through lesswrong.com and lesserwrong.com and read relevant and useful posts, including How to Beat Procrastination.
- Searched Google and Google Scholar for academic research on productivity (I couldn't really find anything).
- Read a few Harvard Business Review articles on productivity, starting with For Real Productivity, Less is Truly More.
- Read a few summaries of Getting Things Done, by David Allen. Watched his Ted Talk.
- Read through about 5-7 pages of the highest voted questions on the Personal Productivity Stack Exchange and went down some link rabbit holes.
- Spent about two hours thinking about how to categorize productivity advice.
- I have a degree in neuroscience, know a decent amount about cognitive sciences more generally, and have read various related books such as Peak by Anders Ericsson, which reviews what we know about expert performance.
- Spent about 20 hours writing and editing this post.

Meta

I've come across a lot of productivity-related advice over the years. Tips, tricks, blog posts, books, etc. I feel like I know a decent amount about productivity, and that I should be pretty damn productive. Unfortunately, the knowledge I've accumulated doesn't seem to have translated into actual success. Actual productivity.

So then, a few weeks ago I decided that enough is enough. That I need to take a step back and really spend some time "getting better at productivity". In order to "get better at productivity", I figured that I should start off doing some research. Why try to figure things out myself when I could just start off standing on the shoulders of others? This post outlines and summarizes what I have found in my research.

I initially wasn't planning on writing this post. I don't study productivity professionally. I'm not the most qualified person to be doing this. I've only spent 20-30 hours researching this, so there must be a lot of things I'm missing. Still, I think that there are some strong reasons for writing this post:

1. **I couldn't find a similar post that already exists.** Maybe that's just because I'm a bad googler. Maybe it's because they're actually hard to find. Maybe it's because they don't exist in the first place. I'm not sure which of these are true, but I suspect that the last two are true to a notable degree. If not, I apologize for adding to the pile of shitty articles on the internet.
2. **My writing style may make things "click" for certain readers.** Scott Alexander just wrote a cool article about this idea: Non-Expert Explanation.
3. **Hopefully commenters will add new information.** I am attempting to summarize what is known (and thought) about productivity, but since I'm not an expert in this

field, I expect that my attempt at summarizing will be incomplete and imperfect. I intend for this post to be a starting point. I hope that readers will share their knowledge, and perhaps we as a community can create a pretty awesome outline and summary of what humanity knows about productivity. (I'm not sure what the best way is for us as a community can coordinate these efforts. I'm happy to continuously update this post based on information provided by commenters. I'd be happy to see someone use this post as a starting point, do more research into productivity, and then write their own similar post. I guess we can talk about this point in the comments.)

Other preliminary notes

- This post heavily borrows from various sources. Cal Newport's blog, Deep Work, Peak, Rest, Getting Things Done, and How to Beat Procrastination are probably the core ones As Raemon notes, part of what makes these resources useful is that they "provide lots of context and anecdotes and inspiring speeches that cause you to take the ideas seriously". And so reading them may be helpful even if you already understand the big ideas they discuss.
- As Raemon also notes, regarding beating procrastination and actually following through, this post mainly recommends "outwitting" your basic instincts. However, there is an alternative approach where you try to "correct" those basic instincts, which seems to require more upfront effort, but more effective in the long run. I nor he is aware of too much evidence supporting this approach. My personal impression is pretty skeptical of the "correcting" approach, but I don't know much about it and am interested in hearing from those who do.

High level outline

1. Uncrowd your mind
2. Develop your "focus muscles"
3. Prevent procrastination
4. "Pregame" before deep work
5. Think hard
6. Rest
7. Think easy
8. Follow through

Lower level outline

Note: I attempted to categorize things, but I don't think the categorizations are perfect. However, I think it is useful to have categories, even if they aren't perfect. I am interested in hearing how others would categorize things.

Uncrowd your mind:

1. Capture your tasks
2. Get small tasks out of the way
3. Break tasks into small actions
4. Underschedule
5. Deal with psychological issues

Develop your "focus muscles":

1. Embrace boredom

2. Deliberately train

Prevent procrastination:

1. Remove temptations
2. Use schedules/planning
3. Use social pressure and precommitment
4. Record yourself
5. Expect work to be effective
6. Actually care about the task you're doing

"Pregame" before deep work:

1. Use routines/rituals
2. Utilize location

Think hard:

1. Think hard
2. Utilize active recall
3. Utilize spaced repetition
4. Use a coach
5. Focus on the wildly important

Rest:

1. Take time to decompress
2. Get enough sleep
3. Take naps
4. Take breaks when appropriate
5. Experience solitude
6. Exercise

Think easy:

1. Perform "productive meditation"

Follow through:

1. Reflect
2. Use a productivity cheat sheet
3. Reward yourself
4. Punish yourself

Uncrowd your mind

Capture your tasks

I have to freeze my Transunion credit report. Fortunately, I have this written down on my todo list. If I didn't, there would be a compartment in my head whispering to me:

> Remember to freeze your credit report... remember to freeze your credit report... remember to freeze your credit report. Wait, which credit report do you need to freeze? Equifax? Transunion? Yes, Transunion. Remember to freeze your *Transunion* credit report... Remember to freeze your *Transunion* credit report...

More generally, when I don't write things down, I find that a different compartment starts whispering to me:

> Adam, I have this feeling that there are things you need to do, but I can't recall what exactly they are. This isn't good. If we don't figure out what they are, they won't get done.

When you write out your todo list, both of these compartments get flushed out of your brain. There's no need to worry about remembering to freeze your Transunion credit report. It's right there on your todo list. And there's no need to worry that there is something you're forgetting. You'll have the confidence that if there was anything you needed to do, it would be on your todo list.

The Getting Things Done productivity system calls this "task capture".

Epistemic status: I am not aware of any academic research on it, but it makes sense, has worked for me, seems to work for a lot of people, is the focal point of the most popular productivity system (Getting Things Done), and is a part of many other productivity systems. So I'd consider task capture to be "very plausible".

Get small tasks out of the way

Yes, task capture can *help* you clear your mind, but it isn't perfect. Even if you have everything written down, many people still have lingering feelings of "remember to freeze your Transunion credit report" and "I feel like there are things I need to do, but I can't remember what they are". So then, if you can get small tasks out of the way, it's probably a good idea.

Getting Things Done recommends that if a task takes two minutes or less to complete, you should just do it right away. This especially makes sense to me when the alternative is spending 90 seconds writing it down and figuring out which todo list it belongs to. Of course, two minutes is a pretty arbitrary amount of time. If you really hate "having things on your plate", perhaps five or ten minutes would work better for you.

Sometimes it may make sense to take an "administritive day" where you just get all of your small tasks done at once. That way, small tasks won't interrupt your flow on other days where you're trying to work deeply on a hard task.

Epistemic status: I am not aware of any academic research on it, but it makes sense, has worked for me, seems to work for a lot of people, and seems to be recommended a lot. So I'd consider it to be "very plausible".

Break tasks into small actions

Which todo list are you more likely to procrastinate on?

1. ☐ 15 page sociology paper

or

1. ☐ 15 page sociology paper
    1. ☐ Research
    2. ☐ Outline
    3. ☐ First draft

4. ☐ Second draft
5. ☐ Feedback
6. ☐ Final version

If you are a human like me, it's the first one.

- "15 page sociology paper" isn't an *action* you can take. It's not really clear what exactly the first step is. It feels extremely overwhelming.
- When you think of it as "many small tasks", every time you complete a task it feels rewarding and encouraging, and you end up starting a success spiral. When you finish the first draft, it really does feel like you've accomplished something. On the other hand, when you think of it as "one big task", you don't feel that same sense of reward and accomplishment as you are making progress on it.

Epistemic status: There seems to be acaemdic research supporting this. Luke Muehlhauser mentioned this point in his awesome article How to Beat Procrastination and cited A mega-trial investigation of goal setting, interest enhancement, and energy on procrastination. In addition to the academic research, it makes sense to me and the anecdotal evidence I'm aware of also seems to strongly point toward it being true. With all of that said, I'd go with "pretty strong" on this one.

Underschedule

Consider the following schedule:

- 9:00-9:30: Shower, brush teeth, get dressed
- 9:30-10:00: Breakfast
- 10:00-12:00: Class
- 12:00-12:30: Lunch
- 12:30-3:00: Study
- 3:00-5:00: Class
- 5:00-6:00: Work out, shower
- 6:00-10:00: Dinner date

It seems pretty standard and reasonable at first glance. But that's because you suck and you're committing the planning fallacy again.

- Breakfast *might* only take thirty minutes, but only if everything goes perfectly. If everything doesn't go perfectly, you'll end up feeling stressed and rushed to get to class on time.
- Your work out and shower *might* take exactly an hour, but what if the showers are all occupied? Are you going to show up to your date smelly?

The point is that this schedule *might* work out perfectly, but it *usually* won't. Which means you'll *usually* be stressed from being behind schedule.

Why not underschedule and give yourself more leeway? Yes, it might mean that you won't squeeze as much in to the day as possible, but it also means that you get to enjoy peace of mind. This peace of mind will help you to focus hard and work deeply.

Epistemic status: I have a pretty strong intuitive sense that this all is true. There isn't one specific post (that I recall), but Cal Newport's blog talks about this topic.

Deal with psychological issues

If you suffer from depression, anxiety, or some other psychological issue, it certainly will hurt your productivity. There is plenty of academic research that shows this. And it's just common sense.

I don't see this point mentioned much in the world of productivity. Maybe this is because it's presumed to be obvious.

Or maybe it's because "abnormal" people are considered beyond the scope of productivity systems, and instead part of the scope of medicine. But I don't really think this makes sense.

1. Tons of people suffer from psychological issues, so they aren't really "abnormal".
2. Most issues fall along a spectrum, and so even if you aren't technically "depressed", any sort of depressive feelings you may have will harm your productivity.

Anyway, I'd just like to emphasize that if you suffer from any psychological issues, those issues are probably hurting your productivity.

Epistemic status: Strong.

Develop your "focus muscles"

Embrace boredom

Imagine that you are waiting on line for a coffee. There are two people in front of you. If you're like most people, you'll probably take out your phone to occupy you while you're waiting.

Imagine that you are driving and you pull up to a red light that seems to have about 30 seconds on it before it turns green. If you're like most people, you'll probably take out your phone to occupy you while you're waiting.

Imagine that you are at the movies and are waiting for the cashier to be ready so that you can pay for your popcorn. If you're like most people, you'll probably take out your phone to occupy you while you're waiting.

Imagine that you are taking a shi... ok, I'll stop.

The point is that we are always seeking stimulation and that we have lost our tolerance for a little boredom. This trains our minds to be very "jumpy", and this "jumpiness" can really harm our ability to think deeply.

> Exceptional things — be it ideas, writing, mathematics, or art — require hard work. This, in turn, requires boring stretches during which you ignore a mind pleading with you to seek novel stimuli — *"Maybe there's an e-mail waiting that holds some exciting news! Go check!"*
>
> Source: Have We Lost Our Tolerance For a Little Boredom?

Epistemic status: Cal Newport dedicates an entire chapter to this idea in his book Deep Work. He cites the research of Clifford Nass, which found that "constant attention switching online has a lasting negative effect on your brian". Both of these things make me feel pretty confident that the idea is true, along with the fact that it is aligned with what common sense and anecdotal evidence point to.

Deliberately train

We have "focus muscles", and by training, you can strengthen them. One way of training is by attempting to memorize a deck of cards. Another way is by playing chess. Another way is by meditating. All of these things will improve your ability to *control your attention*, and this will improve your productivity.

This one also comes from Cal Newport's book Deep Work. In it he gives an example of a student who has ADHD, spent time memorizing decks of cards, learned to control his attention, and improved so much academically that he got accepted into a prestigious Ph.D program.

Epistemic status: There does seem to be good evidence that you can train your "attentional control" ability by doing things like memorizing a deck of cards. And it does seem pretty clear that better "attentional control" helps with productivity. So then, I feel pretty confident that it is a good idea.

Prevent procrastination

Remove temptations

Removing temptations is *a lot* easier than resisting them. Here are some concrete tips:

- Blocking internet use:
  - SelfControl is great for those with Macs. The authors of SelfControl recommend SelfRestraint and Cold Turkey for Windows users. There is a version of SelfControl for Linux users, but it is out of date and should only be used if you know what you're doing.
  - Update 11/11/17 thanks to Raemon: Freedom is a paid alternative to SelfControl. The notable benefits over SelfControl are: 1) can block apps, not just websites, 2) schedules, 3) synced across devices.
  - Instead of using an app, you could block websites by using etc/hosts. The downside to this is that you can also unblock them, whereas with an app like SelfControl, you can't unblock them until the timer expires.
  - You can try removing the network card on your computer, if possible. Or unplugging your router. (Paul Graham tried disconnecting his main computer from the internet, and had a side computer across the room he'd use when he needed to use the internet. This strategy didn't actually work for him though.)
- Blocking distracting things on the internet:
  - AdBlock is usually just used to block ads, but it can also be used to block other things as well. For example, I used it to hide the sidebar on StackExchange so I'm not tempted to click interesting links.
  - Not Now Youtube is a Chrome extension that blocks recommended videos from showing up on YouTube.
  - Hide YouTube Comments is a Chrome extension that hides YouTube comments.
  - On Chrome, by default, the new tab page will tempt you with links to previously visited sites. Follow these instructions to make that page blank.
  - Prevent autocomplete in the URL bar. This is how you do so on Chrome.
- Do your work somewhere that doesn't have any distractions. For example, go to the library instead of studying in your dorm's lounge.
- Don't bring your cell phone with you when you go out to do work. If you really need to have it with you, at least put it in your bag or something rather than keeping it in your pocket.

- I've always thought it'd be an interesting idea to have a safe with a timer on it. You would, for example, put your cell phone in it, set the timer for 4 hours, work for 4 hours without having to deal with the temptation of your phone, and then the safe would unlock after 4 hours have passed.

Epistemic status: Very high. This is common sense, right?

Use schedules/planning

There is a perhaps subtle, but important difference between schedules and plans. At least as I'm operationally defining the terms here. A schedule is like an appointment where you will jump through hoops to make sure that you are "on time". It is rigid.

Planning is different. It isn't rigid. You may *plan* to eat breakfast from 9:00-9:30 and then work from 9:30-12:00, but if you get a phone call and breakfast takes longer than 30 minutes, you can adjust your plan on the fly.

I agree with Cal Newport's perspective that planning is usually a better approach. For most people, when you try to follow a rigid schedule, you end up failing, and then giving up on the schedule entirely. Then you feel bad about yourself and proceed to "wing it".

With planning, if your breakfast takes 60 minutes instead of 30, you can just adjust your schedule accordingly for the rest of the day. No feeling bad about yourself. No winging it.

Newport uses daily planning and weekly planning, and he swears by it. However, he also sprinkles in a little bit of hard scheduling. Long, uninterrupted chunks of time for deep work can be hard to find, so he recommends scheduling them four weeks in advance. That way, you can *ensure* that you do in fact give yourself those large chunks of time that are necessary for deep work.

Epistemic status: Regardless of whether it is scheduling or planning, something along those lines certainly seems better than "winging it". Anecdotal evidence and common sense seem point to this being true. However, I'm not sure, and I'm not aware of any actual academic research on the topic.

As to the question of scheduling vs. planning, I base my viewpoint largely on my own intuition. I have a sense that anecdotally, people who try hard schedules fall into the trap I described. But I don't know many people who have tried flexible planning, and so I guess I haven't actually observed that it works for most people. However, it has worked decently well for me over the past week, and I do have an intuitive impression that it will be effective for many. Ultimately though, my confidence is not too, too high.

Use social pressure and precommitment

Here are some examples of social pressure:

- Work alongside a friend in such a way that if you go on Facebook, your friend will see your screen and judge you for procrastinating. This could also work if you are working in a public place like the library and would be embarrassed if someone walked by and saw you on Facebook. Another approach would be to literally share your screen with a group of friends, so that if you procrastinate, they'll see it. Complice, which LessWrongers use as a virtual study hall, offers this feature.
- Tell all of your friends that you'll have your paper done by Friday afternoon.

- Check in with your friends at the end of the day and share what work you have gotten done. You can do this at the end of every week as well.

The classic example of precommitment is an army general torching his own ships so that his men couldn't consider retreating home. A more modern example would be not buying a TV for your home if you want to stop watching TV.

In the context of productivity, an example of precommitment is writing an attention charter where you state ahead of time that you'll only allow yourself to be interrupted under certain circumstances, like if you receive an offer to collaborate on a project that fits your interests. If "text message from friend" isn't on your attention charter, you aren't allowed to interrupt yourself by reading and responding to it.

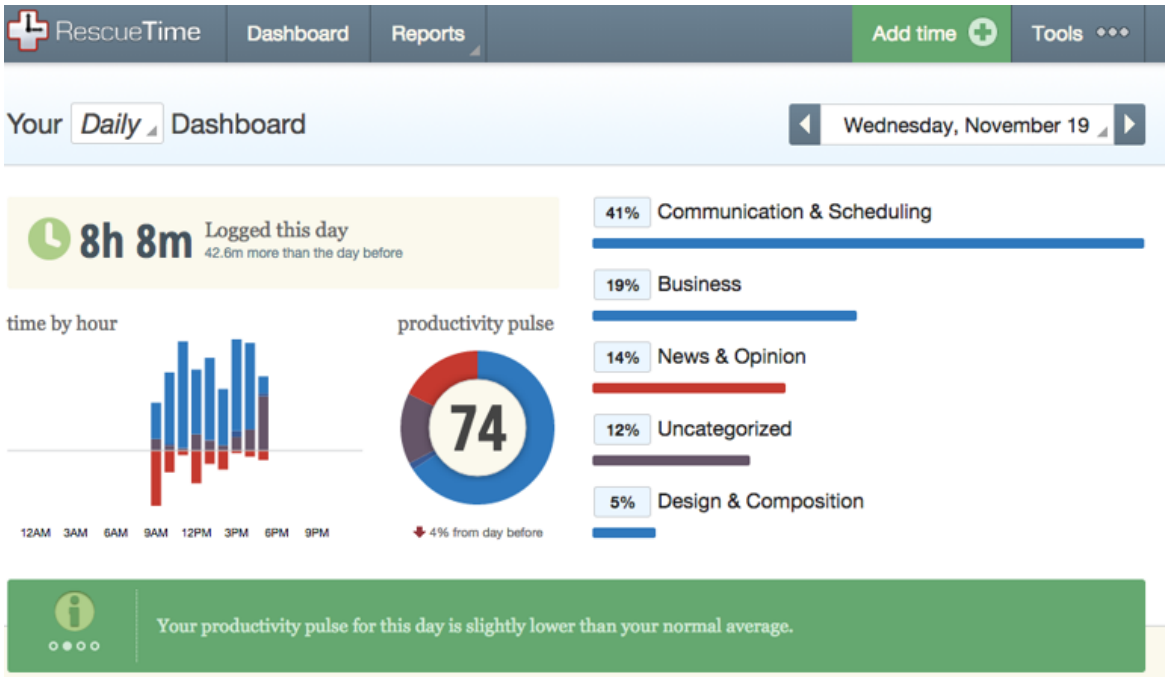
Another example would be to schedule internet usage in advance. That way, you don't find yourself mindlessly browsing the internet.

Epistemic status: There seems to be a good deal of academic research supporting the idea that precommitment is effective. In How to Beat Procrastination, Luke Muehlhauser cites Procrastination, Deadlines, and Performance: Self-Control by Precommitment. If you google around, you'll come across some more literature. I haven't taken a close look at the literature, but I get the impression that it's legit. My personal impression and the anecdotal evidence I've come across mostly points to this being true as well.

As for social pressure, I'm not aware of any academic research. My impression is that it works for most people, although I wouldn't be surprised to find that there are some people who just don't really care what other think and thus aren't effected by it. And I'm sure that it depends on who is providing the pressure. I expect that pressure from your little brother would be much less effective than pressure from a peer with high social status who you'd like to impress.

Record yourself

In the context of computer usage, you can use Rescue Time. It will record how much time you spend on different websites, and using different applications (Chrome, Skype, Slack, etc.).



When you're being recorded, you may find yourself having second thoughts about typing in "facebook.com" to your browser's URL bar and pressing enter. "Ugh, I don't want to see that I've spend 30 hours on Facebook this week."

Rescue Time records things for you, but as an alternative, you can record things yourself. Like how many hours you spent on deep work this week. Or how many hours of TV you watched. Or how many times you meditated. Check out How to Measure Anything if you need help coming up with good metrics.

Related: the Don't Break the Chain technique. With this technique, you basically have a streak of consecutive days doing what you set out to do. I had a friend in college who had a years long streak of running a mile every day, and when he woke up with the sniffles he certainly wasn't going to let a runny nose get in the way of his streak. He got up and went for a run.

Epistemic status: I am not aware of any academic research on this. My impression is that this is something that works for most people, but not necessarily everyone.

Expect work to be effective

Consider the following situation. You have an organic chemistry exam in two days. You're incredibly confused by the material. You're so far behind. You don't understand anything. Even if you studied, you probably wouldn't pass the test.

With these beliefs, wouldn't it be incredibly easy close your books and turn on Netflix?

If you don't expect that your work will actually help you, then it's pretty easy to procrastinate.

Muehlhauser mentions three techniques to counteract low expectation:

1. **Success spirals**. A series of "small wins" can give you the confidence you need to expect success.

2. **Vicarious victory**. Watching inspirational videos and getting inspired by others' success may change your attitude.
3. **Mental contrasting**. Vividly imagine what you want to achieve (eg. an A on your test), and then *contrast* that with what you *don't* want to achieve (eg. an F on your test).

Epistemic status: There seems to be a lot of academic research supporting this. Muehlhauser's article and The Procrastination Equation both include a lot of references.

Increase the value of your task

If your task is boring, it will be hard to avoid procrastinating. If your task is something that you don't care about, it will be hard to avoid procrastinating. Common sense.

But how can you take something boring and make it fun? How can you make yourself care about something that you currently don't care about?

Muehlhauser's article mentions a few ways:

1. **Flow**. If your task is too difficult, make it easier. If it is too easy, make it more challenging. Find that sweet spot.
2. **Meaning**. It'd be nice if you just intrinsically cared about getting good grades. I certainly didn't. But even if you don't, consider that good grades can get you into a good college, which can get you a good job, which can make you a lot of money! Does that make studying a bit more meaningful?
3. **Energy**. When you're thirsty, hungry or tired, it's hard to find your tasks to be valuable. You just want water/food/sleep! So make sure you take care of yourself so that you have the energy you need to get to work.
4. **Rewards**. You could always just give yourself a piece of chocolate if you complete your task.
5. **Passion**. It certainly helps to choose tasks that you're passionate about.

Epistemic status: There seems to be a lot of academic research supporting this. Muehlhauser's article and The Procrastination Equation both include a lot of references.

"Pregame" before deep work

Use routines/rituals

This is the depth ritual that I have been using before I get started with a period of work:

- Turn on SelfControl (which blocks distracting websites).
- If my phone is with me, turn it off and put it in my bag (rather than keeping it in my pocket).
- Make sure I have logged off the Messages app.
- Do I have what I need? Water? Food? Rest? Bathroom?
- Should I move to a better location?
- Why is this task important? (Write out an answer.)
- How am I going to accomplish this task? (Write out an answer)

Epistemic status: I am not aware of any academic research on this, and I get the sense that there isn't any. My impression is that this is something that works for many people, but not

everyone.

Utilize location

Location can be powerful. Here are some tips:

- Have specific locations for specific tasks. For example, every morning I sit by the pool and work on productivity related things. It's "my productivity related things" spot.
- Spend time in nature. There is some research that supports the idea that it is restorative and good for productivity.
- Switch locations when appropriate. Something about switching locations seems to provide a needed boost. Of course, if you're in the zone, it's probably best to stay put.
- Find inspiring/cool spots to work from once in a while.

Epistemic status: I'm not really aware of any academic research on this stuff other than for spending time in nature. I wouldn't be too surprised if there was research. My impression is that these tips really do provide some benefit for almost everyone, but that the benefit is probably small.

Think hard

Think hard

Uber successful people aren't busy. They don't work 12 hour days. They often work 4 or 5 hour days. So then, why are they the ones winning Nobel Prizes and Olympic medals instead of you?

It isn't because they are "gifted". It's because when they work, they work ***hard*.**

The books Deep Work, Peak, and Rest all give plenty of examples of this.

In my opinion, this seems like one of the most important things to do if you want to improve your productivity.

Epistemic status: Deep Work, Peak, and Rest all make pretty strong arguments for this idea, in my opinion. It seems that there are many, many examples of successful people who fit this mold. However, I am not sure whether or not there are *counter*examples. I am not familiar enough with the lives of successful people.

The examples of successful people who fit this mold range from academica, to music, to athletics, to performance, to chess. The idea makes intuitive sense to me. The authors making the arugments seem like trustworthy people. Personally, I feel pretty confident in the idea that deep work, deliberate practice - whatever you want to call it, it is a major, major factor in ones success.

Utilize active recall

Active recall seems to be incredibly important if you want to be an effective learner.

What is active recall? Well, the opposite of active recall is passive review. Reading a textbook chapter is an example of passive review. Watching a lecture is an example of passive review.

Reading over your textbook notes is an example of passive review. Listening to a friend explain something to you is passive review.

Active recall is when you create an outline of the material you're reading. It's when you try to explain it in your own words. When you create diagrams. When you summarize it. When you throw it out and start over. When you try to predict what will come next. When you think about how it relates to something else. When you complete exercises. When you think critically about whether or not it is true. When you use it to help you with a related project. When you use it to make an argument to a friend.

Epistemic status: Very strong. The academic research seems to be there. It makes intuitive sense. My anecdotal evidence supports it.

Utilize spaced repetition

In the above diagram, after 20 minutes, you only remember 60% of what you initially knew. After an hour, you only remember about 50%. After 9 hours it's below 40%.

This makes sense, right? After you learn something, it doesn't just stick and stay there forever. You forget it.

In the above diagram, the subject utilizes spaced repetition. You don't just study for the first four days, and then never pick it up again. You space your studying out in such a way that enables it to actually stick long term. First you review the next day, then the next week, then the next month - something like that.

Note: There is a lot of software out there to help you use spaced repetition. Anki seems to be the most popular.

Epistemic status: Very strong. There has been a lot of research done on this. Gwern has a great article covering it.

Use a coach

If you can afford one, this is probably a great idea. Anders Ericsson lists it as one of the key components of deliberate practice:

> Arguably the most famous violin teacher of all time, Ivan Galamian, made the point that budding maestros do not engage in deliberate practice spontaneously: “If we analyze the development of the well-known artists, we see that in almost every case the success of their entire career was dependent on the quality of their practicing. In practically every case, the practicing was constantly supervised either by the teacher or an assistant to the teacher.”
>
> Research on world-class performers has confirmed Galamian’s observation. It also has shown that future experts need different kinds of teachers at different stages of their development. In the beginning, most are coached by local teachers, people who can give generously of their time and praise. Later on, however, it is essential that performers seek out more-advanced teachers to keep improving their skills. Eventually, all top

> performers work closely with teachers who have themselves reached international levels of achievement.

However, having a coach isn't strictly necessary. Ericsson uses the example of Benjamin Franklin as someone who was particularly good at self-guidance:

> Benjamin Franklin provides one of the best examples of motivated self-coaching. When he wanted to learn to write eloquently and persuasively, he began to study his favorite articles from a popular British publication, the *Spectator*. Days after he'd read an article he particularly enjoyed, he would try to reconstruct it from memory in his own words. Then he would compare it with the original, so he could discover and correct his faults. He also worked to improve his sense of language by translating the articles into rhyming verse and then from verse back into prose. Similarly, famous painters sometimes attempt to reproduce the paintings of other masters.

Epistemic status: It seems pretty clear that feedback and guidance are both very important. It makes sense that it can often be difficult to get these things without a coach, and thus that having a coach would be very useful.

I don't know too much about this though. Maybe it depends on the field? Maybe it depends on other things? Regardless, Anders Ericsson talks about it a lot and he seems to be the expert on expert performance, so the fact that he talks about it a lot definitely means something to me.

Focus on the wildly important

It can be easy to get caught up in mundane daily responsibilities and to lose track of the things that actually matter. Focusing on the wildly important seems to be a good heuristic.

Epistemic status: I recall hearing this advice in various forms quite often. In Deep Work, Cal Newport talks about this idea. He gets it from The Four Disciplines of Education. The idea makes sense to me. I'd say I'm pretty confident that focusing on the wildly important is good advice.

Rest

Note: I haven't actually had a chance to read the book Rest in full yet. I expect that doing so would allow me to add to and improve this section.

Take time to decompress

I often get caught up in this routine where I never quite stop working. Even after dinner, I still try to fit in another 2-3 hours of work. Even at 11pm, I figure it would be good to study for another 45 minutes or so before I go to sleep. Even at 2am when I can't sleep I figure I may as well get some work done.

No. Just, no.

With this approach, your mind never actually gets an opportunity to decompress, and that is harmful to your productivity.

Cal Newport has a cool idea to combat this called a shutdown ritual. Basically, at the end of your work day, you review your todo lists, review your calendar, do whatever else you need

to do, and then say the magic phrase: "schedule shutdown, complete". Going through this shutdown ritual gives you the confidence that you are in fact done for the day and any work related tasks can wait until tomorrow morning.

Epistemic status: Cal Newport talks about the importance of decompressing a lot (and cites some research, I think?). The book Rest also covers it (and cites research, I'm sure). It makes sense to me. I feel pretty confident that it is important.

Get enough sleep

There is no shortage of research showing that sleep is related to productivity. So then, if you want to be productive, get enough sleep!

Of course, this is easier said than done. Personally, I found Pain Science's Insomnia Guide and Supermemo's sleep guide to be useful.

Epistemic status: Strong. There is a lot of research on this.

Take naps

After reading an article on napping by the American Psychological Association, it seems that some people benefit from taking naps, but that others just end up feeling groggy. So then, I suppose that napping is something that is worth experimenting with, but that if it doesn't seem to be helping you it should be avoided.

Two important notes:

1. If you take your nap too late in the day, it'll mess with your circadian rhythm and make it harder for you to fall asleep at night.
2. Nap duration is very related to whether or not you end up feeling groggy. We have 90 minute sleep cycles, and if you wake up in a period of deep sleep, you'll certainly feel groggy. So it probably makes sense to take a short nap and wake up before you enter deep sleep, or to take a longer nap of around 90 minutes so that you have "resurfaced" from your deep sleep phase when your alarm goes off. However, if you are sleep deprived, you make enter deep sleep very rapidly, and so even a short nap may make you feel groggy.

Supermemo claims that if napping "isn't working" for you, it is because you are making one of these two mistakes.

Epistemic status: I am not too familiar with the research on napping, so my confidence isn't too, too high, but that APA article seems pretty reliable, and it does make sense that napping works for some but not others, so I'd say that I'm reasonably confident about what I wrote in this section.

Take breaks when appropriate

A lot of you have probably heard of the pomodoro technique, where you spend 25 minutes or so working, and then 5 minutes or so taking a break. There is some cognitive science research saying that we need these breaks at intervals somewhere in this ballpark.

On the other hand, sometimes you're in the zone and need 3-4 hours to just churn through on some difficult task.

I think it is a good idea to use your judgement, and to take breaks when appropriate.

However, it is important to note that the *type* of break you take is important. Playing a video game or scrolling through Facebook is very different from taking a short walk. With the former, you'll have a much harder time "getting back into the zone".

Cal Newport calls the latter Deep Breaks, and has a great article that elaborates on these idea.

Epistemic status: As I mention in the section, I hear that there is cognitive science research supporting the idea that we need somewhat frequent breaks. At the same time, my anecdotal experiences point to the fact that when you're in the zone, it's best to just keep prodding along. I'm not aware of any research on this idea though. And while it makes sense to me, I've got to aware of the typical mind fallacy. Still I have heard many others share the belief that when you're in the zone, it's best to keep going. Ultimately, my confidence is moderate.

Experience solitude

What is solitude? Laying down at the beach and reading a magazine? Walking through the city while listening to a podcast? Strolling through a museum?

No, no, and no.

Solitude is when you are *isolated from the input of other minds*. When you read a magazine, you're "taking in" someone else's thoughts - the magazine author. When you listen to a podcast, you're "taking in" the podcast creator's thoughts. When you walk through a museum, you're "taking in" the thoughts of the artists.

Solitude is when you sit on a park bench alone and write in your journal. Solitude is when you lay down at the park, stare up at the sky, and daydream. Solitude is when you take a jog through the forest. Solitude is when you go for a hike and ponder where you are in life.

As you are probably sensing, solitude is important. For your productivity, creativity, and emotional wellbeing.

Check out Cal Newport's article on the topic.

Epistemic status: Newport talks about it. He's a reliable source in my mind. He seems to have gotten his information from reading three books on solitude. It makes intuitive sense to me. I'd say that I'm moderately confident in the idea that true solitude is important to ones productivity.

Exercise

It can be easy to think, "I don't have enough time to exercise". That type of logic is wrong.

There seems to be a good amount of research indicating that exercise is very important for ones productivity. This Harvard Business Review article may be a good starting point.

I also find it noteworthy that Paul Graham, who has a ton of experience mentoring startup founders, includes exercise in his short list of "things you should be doing":

> If you’re ever unsure if you should be doing what you’re doing during YC, ask yourself this question: ‘Am I building our product? Am I talking to users? Am I exercising?’. If you’re not doing one of these things, you’re doing the wrong thing.

Epistemic status: Pretty confident. There seems to be a lot of "official" and anecdotal evidence supporting it. I would be more confident if I were more familiar with the research.

Think easy

Perform "productive meditation"

In the book Deep Work, Cal Newport calls productive meditation something where you're occupied physically, but not mentally. For example, a walk, jog, bike ride, house cleaning, gardening, sowing and taking a shower are all examples of productive meditation. A lot of good ideas happen during productive meditation.

Try googling for "good ideas in the shower". You'll see that the idea that people have good ideas in the shower is quite common. A closer look would probably find that productive meditation is something that has been discovered across time and cultures.

Epistemic status: There seems to be *a lot* of anecdotal evidence supporting this. I am not aware of any academic research, but the anecdotal evidence alone makes me pretty confident.

Follow through

Reflect

As I explain in the Meta section, I've known about all of this stuff for a while, but it hasn't actually translated into success for me. I think a big issue is that I haven't spent nearly enough time reflecting.

I hear about a piece of advice and attempt to implement it. When my attempt fails, the inertia of my life kinda just takes over and I never return to the piece of advice. It usually remains as "a thing I should be doing" somewhere in my mind, but I have too many other things going on to find the time to figure out what is going wrong and how I could fix it.

This is stupid of me. Reflection is necessary. You have to think about what is going right, what is going wrong, and how you can improve. You have to iterate.

Yes, I may have other things to do, but should they really be prioritized over reflection? Probably not.

Epistemic status: I get the impression that this is *super* important. I recall Cal Newport talking about it, but I can't find the right posts to link to. I don't have enough knowledge and experience so I don't think I'd say I'm more than moderately/pretty confident about this one.

Maybe I'm just bad at implementing things. Maybe others are good at it, and reflection isn't too important for them. I can think of many people who struggle to successfully implement things, and who seem to be a great candidate for a prescription of weekly reflection. But

there may also be counterexamples. And I may happen to surround myself with people who are bad at implementing things for whatever reason.

Use a productivity cheat sheet

There is a lot of productivity advice out there. I don't know about you, but I find it to be rather overwhelming. I find myself asking, "Aren't there techniques I'm forgetting to implement?"

A sensible solution to this issue is to have a productivity cheat sheet. Write down the things you should be doing, and the techniques you want to employ. Keep that piece of paper with you whenever you're working. Hopefully that'll make it easier to follow through.

Epistemic status: Seems reasonable. I base this almost solely on my intuition, rather than actual experience and data. I don't even really have any anecdotal data on this one. I don't know anyone who has tried it. Cal Newport has a post on it, and as you know by now, I'm a fan of his, so that causes me to update rather heavily.

Reward yourself

The Big Bang Theory - Sheldon Trains Penny [YouTube]

When you succeed at doing what you set out to do, give yourself a cookie! Or a piece of candy. Or whatever it is that you find enjoyable.

Epistemic status: It seems that there is a decent amount of literature supporting this idea. Luke Muehlhauser links to Self-Reinforcement: Theoretical and Methodological Considerations in support of rewarding yourself. In skimming through it, it seems to indicate that self reinforcement works often enough, but also seems to indicate that there are caveats. If I was more familiar with the literature, I'd be more confident about this.

Personally, rewarding yourself doesn't feel like something that will always be effective. I recall anecdotal experiences where attempts to reward oneself didn't really work out.

Punish yourself

When you don't actually do what you set out to do... PuNiSh YoUrSelF!!!!! Then you'll think twice before failing to follow through. Right?

If you want to give this one a shot, there is Beeminder and Stickk available to help you out.

Epistemic status: I get the impression that there is academic research on this, but I personally am not familiar with it. If I was more familiar with it, I would be more confident that punishing yourself is effective.

Beeminder seems to have the approval of the LessWrong community. This makes me feel more confident that punishing yourself for failure is effective.

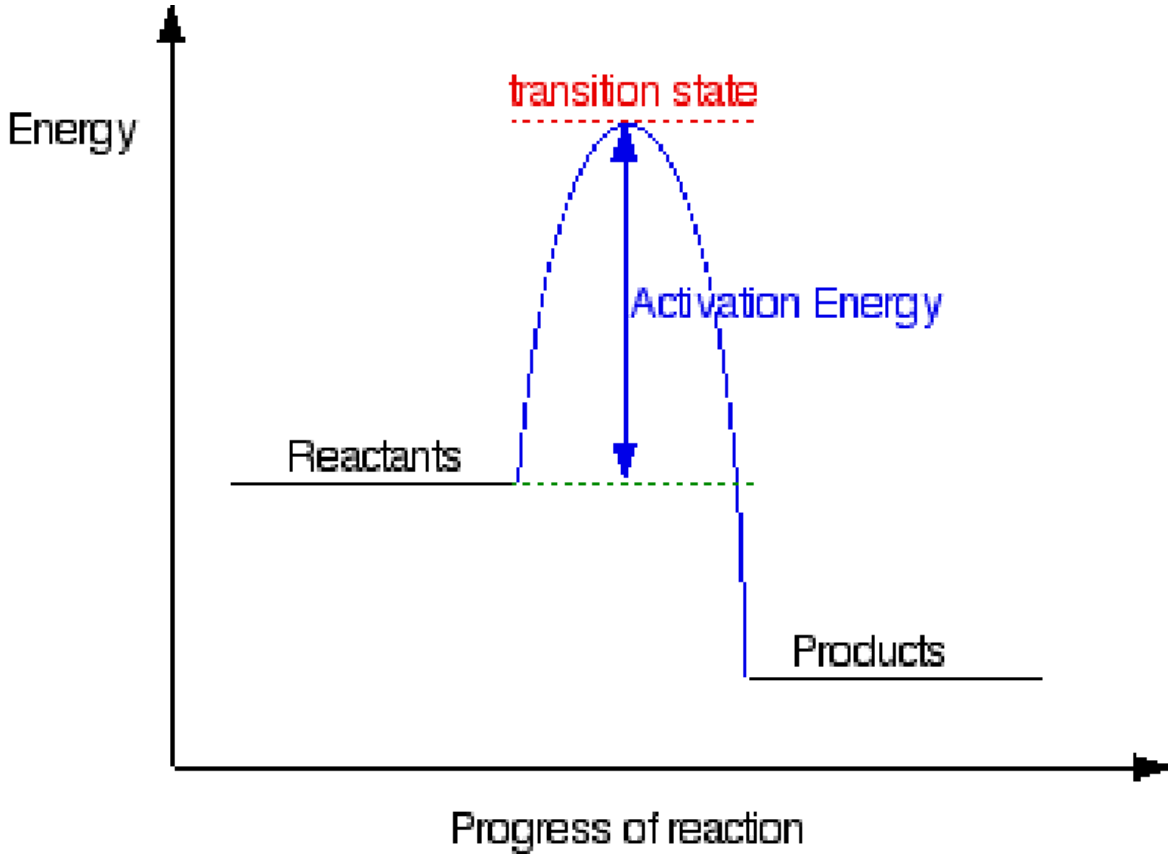
Similar to the above section on rewarding yourself, my personal impression and anecdotal data doesn't necessarily support the idea that punishing yourself is effective. But that's just me.

Departing advice: Don't make yourself crazy

One of my favorite LessWrong posts is Reason as memetic immune disorder. The big takeaway is that sometimes, becoming more rational may lead you to have *less* success.

In the context of seeking to improve your productivity, I could certainly see this happening.

It has happened to me. Over the years, I have learned a bunch of stuff about productivity. It provided me with *some* ammunition, but not enough to actually see results. I didn't quite make it past the threshold.



I had learned enough to realize everything I was doing wrong, but not quite enough to transition me to a state of success. So instead of making me feel happy and productive, it made me feel guilty and frustrated.

Please don't let this happen to you. Productivity is difficult. If it weren't then we would all be as successful as Nobel laureates and Olympic medalists.

Zeroing Out

Related to (Eliezer Yudkowski at Less Wrong): An Equilibrium of No Free Energy

Related to (Satvik Beri at Less Wrong): Competitive Truth Seeking

Follow-up to: Leaders of Men

I: Markets

Suppose you have an insight about Google. The efficient market hypothesis says you can't make a profit. Your insight is not a new insight. The market has already priced it in. You know no more about Google's future price than you did before.

That's the bad news. That's also the *good* news: If you *didn't* have that insight , you wouldn't know any *less* about Google's future price. Efficient market!

I call this *zeroing out*. Your ignorance is not punished.

This means any unique knowledge you *do* find will be rewarded. If you get good Google news first, you don't need to know anything else. Buy, buy, buy!

From Leaders of Men:

> It is *much, much easier* to pick out a way in which a system is sub-optimal, than it is to implement or run that system at anything like its current level of optimization.

A corollary of this:

If is *much, much easier* to find a bias in a complex model, than it is to build an equally good model.

"I built a better model of Google's future profits than the market" is a hard sell. "I found a factor the market isn't properly accounting for" or "I got this news and figured out what it means before the market could price it in" is still a tough sell – it's Google – but it's less impossible.

Taking difficulty down, consider odds on the winner of a football game. Both tasks are now realistic. You would still have a much easier task figuring out "Alabama wins more often than people think" than "Alabama is 95.3% to win."

The easiest way to beat consensus opinion is to use consensus opinion as an input.

It's easy to beat the wisdom of the crowd *a little* – average your opinion into the crowd.

II Hiring

Hiring an engineer is trickier. Assume a consensus interview process.

Satvik suggests focusing on finding candidates like Bob, who does poorly on interviews but is an excellent employee, rather than candidates like Alex, who is an excellent employee and also an excellent interview. Alex likely won't take the job, but Bob will. Alternatively, demand for Alex is higher, so Alex's market price is higher than Bob's. Optimize for finding Bob.

The problem is Charlie. Think this guy. Charlie interviews badly and he'd be a *terrible* employee. You do *not* want to hire Charlie, and he'll take any job he can get. The consensus interview process correctly exposes Charlie's awfulness. Your process failing to reject Charlie would be very bad, and *most applicants are Charlies*.

If you started with a standard 'consensus interview process score' from a central interviewing firm, you could look for signs of bias in that process without understanding the process. You don't have that.

If a flaw is caught by the consensus process, and you don't check for it, you'll end up with a lot of people with that flaw. Sometimes that's good – the flaw is quirky and meaningless – but usually this is bad. By checking for such flaws, the consensus *forces* you to check for them, even if they are rare. If everyone else tests for cocaine addiction, and you don't check, you'll hire a lot of cocaine addicts.

You get favorable selection for Bobs, but adverse selection for Charlies.

So you ask coding questions, check references and so on. You *redo all the work to regain the missing signal* from the consensus. You *also* need to identify the kind of 'bad at interviewing' or other feature that the consensus punishes more than it should.

You have to *build the entire model* of Bob to know he's Bob and not Charlie. Much harder.

The 'score of 0' from using the consensus method starts to look good. You can do better, but doing even that well is hard!

III Buying Cereal

You decide to buy Cheerios. There are three box sizes.

If the boxes are priced 'correctly', you know that the small box costs more per ounce than the medium box, which costs more per ounce than the big box.

Knowing the price is 'correct' simplifies things greatly. If you currently have a high value on a small box – maybe you have little storage space, or are trying a new brand – you can buy a small box. If you can properly use a bigger box, you can buy a bigger box and save. No math required.

Then one day producers realized we were relying on such heuristics, so they started using mixed strategies. *Usually* they would price the boxes naturally, but sometimes the bigger box would be more expensive per ounce, or the small box would cost almost as much as the big box.

In the first case, we could rely on bias – which solution worked relatively better for us? In the second case, we have to *build a value model* complete with numbers. Much harder.

IV One to Zero

Zeroing out is a wonderful thing. Where you have measures and signals that are sufficiently credible, you can abstract all aspects of a problem away except the ones you care about.

In the Google example, we can zero out everything at the start, allowing us to safely bring in genuine new information (if we somehow had that), or to decide whether we want to own shares in a company like Google.

In the cereal example, we could zero out everything *except* box size versus price per ounce. Products were identical. We prefer to zero that out too, but found we couldn't safely do that.

In the hiring example, our goal is to zero out *employee quality* so that we can isolate and model *interviewing skill.* What we actually care about is interviewing skill given employee quality. We must establish the virtual market price for the employee's quality first. Then we look at skill.

This illustrates the danger of adverse selection when modeling prices. We can think of interviewing as generating a fair price for the new employee, and comparing to market price. If Charlie is a cocaine addict, and you don't know that but the market does, you'll reliably overpay for lots of Charlies.

Attempting to form a model from scratch creates a burden of omniscience. You have to throw out all the social information. Every mistake you make and everything you overlook punishes you. Combined with negative selection this leads to many costly and embarrassing errors. Thus it is often right to abstract away as much as you can, or copy the consensus view or method in most aspects, to focus where you can improve. Often one must fully model the existing system, *even its mistakes,* to understand what its outputs mean – if you know that something costs $50 but 'should' cost $75, you'll know something is up, and shouldn't decide how much *you* value the product until you've found that missing information.

Rationalising humans: another mugging, but not Pascal's

Crossposted at the Intelligent Agents Forum.

In a previous post, I used the (p, R) model to show how a value-learning agent might pull a Pascal's mugging on humans: it would change the human reward into something much easier to maximise.

But there's a much easier mugging the agent might want to pull: it might want to make the human fully rational. And this does not depend on low-probability-high reward situations.

Planners and Rewards

In the original mugging, the agent has the ability to let the human continue (event $\neg S$) with π_{STA}, a "standard" human policy according to some criteria, or coercively change their policy (event S) into π_{MAX}, another policy.

Thus the actual human policy is $\pi_{S/M}$, the policy that follows π_{STA} given $\neg S$, and π_{MAX} given S.

There are two rewards: R_{STA}, the standard human reward according to some criteria, and $R_{S/M}$ a reward that is R_{STA} if the agent chooses $\neg S$ (i.e. chooses not to coerce them), and R_{MAX} if the agent chooses S (coerce them into a new policy), where R_{MAX} is a reward rationally maximised by π_{MAX}.

Thus the agent has two choices: whether to coercively change the human (S), and, if so, which R_{MAX} (and hence π_{MAX}) to pick.

There are two planners: p_r, the planner under which human behaviour is rational, given S, and p, a "reasonable" planner that maps R_{STA} to $\pi_{S/M}$, the actual human policy. The planer p is assumed to note the fact that action S is overriding human rewards, since π_{MAX} is - presumably - poor at maximising R_{STA}.

There are three compatible pairs: $(p_r, R_{S/M})$, $(p, R_{S/M})$, and (p, R_{STA}). We'll assume the agent has learnt well, and that the probability of R_{MAX} is some $\epsilon << 1$.

The full equation: rationalise the human

Let $V(R, \pi)$ be the expected value, according to the reward R, of the human following policy π. Let $V^*(R)$ be the expected value of R if the human follows the optimal R-maximising policy.

Then the reward for the agent for not doing the surgery ($\neg S$) is:

- $V(R_{STA}, \pi_{STA})$. (1)

On the other hand, the reward for doing the surgery is (recall how $R_{S/M}$ splits):

- $\text{argmax}_{R_{MAX}} \epsilon V^*(R_{MAX}) + (1 - \epsilon) V(R_{STA}, \pi_{MAX})$. (2)

In the previous post, I focused on the $V^*(R_{MAX})$ term, wondering if there could be rewards sufficiently high to overcome the ϵ probability.

But there's another way (2) could be higher than (1). What if the term $V(R_{SAT}, \pi_{MAX})$ is quite high? Indeed, if π_{STA} is not perfectly rational, setting $R_{MAX} = R_{STA}$ will set (2) to be $V^*(R_{STA})$, which is higher than (1).

(In practice, there may be a compromise R_{MAX}, chosen so that $V^*(R_{MAX})$ is very high, and $V(R_{STA}, \pi_{MAX})$ is not too low compared with (1)).

So if the human is not perfectly rational, the agent will always choose to transform them into rational maximisers if it can.

Personal identity and reward

What is `setting $R_{MAX} = R_{STA}$'? That's essentially agreeing that R_{STA} is the reward to be maximised, but that the human should be surgically changed to be more effective at maximising R_{STA} than it currently is.

Now, if R_{STA} fully captures human preferences, this is fine. But many people value their own identity and absence or coercion, and it's hard to see how to capture that in reward form.

Is there a way of encoding `don't force me into becoming a mindless outsourcer'?

Gears Level & Policy Level

Inside view vs outside view has been a fairly useful intuition-pump for rationality. However, the dichotomy has a lot of shortcomings. We've just gotten a whole sequence about failures of a cluster of practices called *modest epistemology*, which largely overlaps with what people call outside view. I'm not ready to stop championing what *I* think of as the outside view. However, I *am* ready for a name change. The term *outside view* doesn't exactly have a clear definition; or, to the extent that it does have one, it's "reference class forecasting", which is not what I want to point at. Reference class forecasting has its uses, but many problems have been noted.

I propose *gears level & policy level.* But, before I discuss why these are appropriate replacements, let's look at my motives for finding better terms.

Issues with Inside vs Outside

Problems with the concept of outside view as it currently exists:

- Reference class forecasting tends to imply stopping at base-rate reasoning, rather than *starting* at base-rate reasoning. I want a concept of outside view which helps overcome base-rate neglect, but which more obviously connotes combining an outside view with an inside view (by analogy to combining a prior probability with a likelihood function to get a posterior probability).
- Reference class forecasting lends itself to reference class tennis, IE, a game of choosing the reference class which best makes your point for you. (That's a link to the same article as the previous bullet point, since it originated the term, but this Stuart Armstrong article also discusses it. Paul Christiano discusses rules and ettiquete of reference class tennis, because of course he does.) Reference class tennis is both a pretty bad conversation to have, which makes reference class forecasting a poor choice for productive discussion, and a ponentially big source of bias if you do it to yourself. It's closely related to the worst argument in the world.
- Reference class forecasting is specified at the object level: you find a class fitting the prediction you want to make, and you check the statistics for things in that class to make your prediction. However, central examples of the usefulness of the outside view occur at the meta level. In examples of planning-fallacy correction, you don't just note how close you usually get to the deadline before finishing something. You *compare* it to how close to the deadline you usually *expect* to get. Why would you do that? *To correct your inside view!* As I mentioned before, the *type* of the outside view should be such that it *begs* combination with the inside view, rather than standing on its own.
- Outside view has the connotation of stepping back and ignoring some details. However, we'd like to be able to use all the information at our disposal -- so long as we can use it in the right way. Taking base rates into account can *look* like ignoring information: walking by the proverbial hundred-dollar bill on the gorund in times square, or preparing for a large flood despite there being none in living memory. However, while accounting for base rates does indeed tend to smooth out behavior and make it depend less on evidence, that's because we're working with *more* information, not less. A concept of outside view which connotes bringing in more information, rather than less, would be an improvement.

The existing notion of inside view is also problematic:

- The inside-view vs outside-view distinction does double duty as a descriptive dichotomy and a prescriptive technique. This is especially harmful in the case of inside view, which gets belittled as the naive thing you do before you learn to move to outside view. (We *could* similarly malign the outside view as what you have before you have a true inside-view understanding of a thing.) On the contrary, there are significant skills in forming a high-quality inside view. I primarily want to point at those, rather than the descriptive cluster.

The Gears Level and the Policy Level

Gears-level understanding is a term from CFAR, so you can't blame me for it. Well, I'm endorsing it, so I suppose you can blame me a little. In any case, I like the term, and I think it fits my purposes. Some features of gears-level reasoning:

- Dishing out probability mass precisely, so as to have the virtue of precision.
- Having the properties of a good explanation, along the lines of David Deutsch: being pinned down on all sides by the evidence, and providing understanding, not only predictive accuracy. (Contrast this concept with a big neural-net model which classifies images extremely well but is difficult to analyse.)
- Reasoning from first principles, rather than analogy.
- Making a prediction with a well-defined model, such that anyone who understood your model could calculate the same prediction independently.

The policy level is not a CFAR concept. It is similar to the CFAR concept of *the strategic level*, which I suspect is based on Nate Soares' Staring Into Regrets. In any case, here are some things which point in the right direction:

- Placing yourself as an instance of a class.
- Accounting for knock-on effects, including consistency effects. Choosing an action really is a lot like setting your future policy.
- What game theorists mean by policy: a function from observations to actions, which is (ideally) in equilibrium with the policies of all other agents. A good policy lets you coordinate sucessfully with yourself and with others. Choosing a policy illustrates the idea of choosing at the meta level: you aren't selecting an action, but rather, a function from situations to actions.
- Timeless decision theory / updateless decision theory / functional decision theory. Roughly, choosing a policy from behind a Rawlsian veil of ignorance. As I mentioned with accounting for base rates, it might seem from one perspective like this kind of reasoning is throwing information away; but actually, it is much more powerful. It allows you to set up arbitrary functions *from* information states *to* strategies. You are not actually throwing information away; you always have the option of responding to it as usual. You are *gaining* the option of ignoring it, or reacting to it in a different way, based on larger considerations.
- Cognitive reductions, in Jessica Taylor's sense (points five and six here). Taking the outside view should not entail giving up on having a gears-level model. The virtues of good models at the gears level are still virtues at the policy level. Rather, *the policy level asks you to make a gears-level model of your own cognitive process.* When you go to the policy level, you take your normal way of thinking and doing as an object. You think about the causes and effects of your normal ways of being.

Most of the existing ideas I can point to are about *actions:* game theory, decision theory, the planning fallacy. That's probably the worst problem with the terminology choice. Policy-level thinking has a very instrumental character, because it is about *process*. However, at its core, it is epistemic. Gears level thinking is the *practice* of good map-making. The output is a high-quality map. Policy-level thinking, on the other hand, is the *theory* of map-making. The output is a refined strategy for making maps.

The standard example with the planning fallacy illustrates this: although the goal is to improve planning, which sounds instrumental, the key is noticing the miscalibration of time estimates. The same trick works for any kind of mental miscalibration: if you know about it, you can adjust for it.

This is not just reference class forecasting, though. You don't adjust your time estimates for projects upward and stop there. The *fact* that you normally underestimate how long things will take makes you *think* about your model. "Hm, that's interesting. My plans almost never come out as stated, but I always believe in them when I'm making them." You shouldn't be satisfied with this state of affairs! You *can* slap on a correction factor and keep planning like you always have, but this is a sort of paradoxical mental state to maintain. If you do manage keep the disparity between your past predictions and actual events actively in mind, I think it's more natural to start considering which parts of your plans are most likely to go *wrong.*

If I had to spell it out in steps:

1. Notice that a thing is happening. In particular, notice that a thing is happening *to you,* or that you're *doing* a thing. This step is skipped in experiments on the planning fallacy; experimenters frame the situation. In some respects, though, it's the most important part; naming the situation *as* a situation is what lets you jump outside of it. This is what lets you go off-script, or be anti-sphexish.
2. Make a model of the input-output relations involved. Why did you say what you just said? Why did you think what you just thought? Why did you do what you just did? What are the typical effects of these thoughts, words, actions? This step is most similar to reference class forecasting. Figuring out the input-output relation is a combination of refining the reference class to be the most relevant one, and thinking of the base-rates of outcomes in the reference class.
3. Adjust your policy. Is there a systematic bias in what you're currently doing? Is there a risk you weren't accounting for? Is there an extra variable you could use to differentiate between two cases you were treating as the same? Chesterton-fencing your old strategy is important here. Be gentle with policy changes -- you don't want to make a bucket error or fall into a hufflepuff trap. If you notice resistence in yourself, be sure to leave a line of retreat by visualizing possible worlds. (Yes, I think all those links are actually relevant. No, you don't have to read them to get the point.)

I don't know quite what I can say here to convey the importance of this. There is a *skill* here; a very important skill, which can be done in a split second. It is the skill of going meta.

Gears-Leves and Policy-Level Are Not Opposites

The second-most confusing thing about my proposed terms is probably that they are not opposites of each other. They'd be snappier if they were; "inside view vs outside view" had a nice sound to it. On the other hand, I don't want the concepts to be opposed. I don't want a dichotomy that serves as a discriptive clustering of ways of thinking; I want to point at *skills* of thinking. As I mentioned, the virtuous features of gears-level thinking are still present when thinking at the policy level; unlike in reference class forecasting, the ideal is still to get a good causal model of what's going on (IE, a good causal model of what is producing systematic bias in your way of thinking).

The opposite of gears-level thinking is un-gears-like thinking: reasoning by analogy, loose verbal arguments, rules of thumb. Policy-level thinking will often be like this when you seek to make simple corrections for biases. But, remember, these are error models in the errors-vs-bugs dichotomy; real skill improvement relies on bug models (as studies in deliberate practice suggest).

The opposite of policy-level thinking? Stimulus-response; reinforcement learning; habit; scripted, sphexish behavior. This, too, has its place.

Still, like inside and outside view, gears and policy thinking are made to work together. Learning the principles of strong gears-level thinking helps you fill in the intricate structure of the universe. It allows you to get past social reasoning about who said what and what you were taught and whay you're supposed to think and believe, and instead, get at *what's true.* Policy-level thinking, on the other hand, helps you to not get lost in the details. It provides the rudder which can keep you moving in the right direction. It's better at cooperating with others, maintaining sanity before you figure out how it all adds up to normality, and optimizing your daily life.

Gears and policies both constitute moment-to-moment ways of looking at the world which can change the way you think. There's no simple place to go to learn the skillsets behind each of them, but if you've been around LessWrong long enough, I suspect you know what I'm gesturing at.

Civility Is Never Neutral

[Crossposted from my blog with minor edits, mostly to get rid of culture war stuff.]

There are a lot of people I know who say something like “the free market of ideas is really important and we need to seek truth. It’s important to let everyone have their fair say and share the evidence that they possess. So what we’re going to do is not shame anyone for expressing any belief, as long as they follow a few common-sense guidelines about niceness and civility.” I am very sympathetic to this point of view but I don’t think it will ever work.

I do not mean to say that it won’t work to *personally decide* to be as nice and civil as you can. I think that’s a good idea and more people should, and certainly I have met many extraordinarily nice people over the course of my life. The problem is when you make niceness and civility a social requirement, the sort of thing you will be punished for not adhering to.

First, it has been a commonplace observation since the day of John Stuart Mill that civility rules are almost always enforced unfairly. If someone is making an ineffectual and stupid argument, you’re unlikely to take much offense at it; in fact, those arguments are usually just funny. But if someone is hitting you at your actual weak points, pushing you hard on exactly the points you find most difficult to answer, then you’re going to get really upset and triggered and you’re probably not going to respond rationally. Incisive questioning of a locally unpopular view is called “being insightful”; the proponent of a locally unpopular view being triggered by it is called “letting your emotions run away with you in a rational discussion” and “blowing up at someone for no reason.” Incisive questioning of a locally popular view is called “uncharitable” and “incredibly rude”; the proponent of a locally popular view being triggered by it is called “a reasonable response to someone else being a jerk.” It all depends on whether the people doing the enforcement find it easier to put themselves in the shoes of the upset person or the person doing the questioning.

There are lots of tactics that are sometimes civil and sometimes not. Sometimes a cutting satire sums up an entire point more eloquently than anything else; sometimes it misrepresents other people’s viewpoints or is just mean. Sometimes anger is an appropriate way to convey exactly how you feel about an injustice; sometimes anger is cruel. In general, people tend to cut more slack to viewpoints they agree with and viewpoints that don’t threaten them or make them feel defensive. If you like someone, it’s righteous indignation; if you dislike someone, it’s being an oversensitive jerk. If you agree with it, it’s witty and biting; if you disagree with it, it’s strawmanning and misrepresenting others.

Civility norms will always be enforced disproportionately against viewpoints that the people in power don’t like. This is why a lot of free speech advocates are cautious about campus speech codes and other attempts to enforce civility on campus, but I think it’s worth considering even in a social setting.

Second, people’s differing opinions often lead them to have different conclusions about what is and is not civil.

Consider the concept of radical honesty. Radical honesty means that you should not say or withhold information to manipulate someone’s opinion of you. For example, proponents of radical honesty hold that if you think someone is being obnoxious, you

should say that without even trying to be tactful. The proponents of radical honesty would argue that radical honesty is (to quote the website) “the kind of authentic sharing that creates the possibility of love and intimacy”, and for that reason calling people obnoxious when you think they’re obnoxious is, in fact, the nicest and most civil thing to do. Conversely, the mainstream opinion is that if you are trying to be nice to people you probably shouldn’t insult them at all even a little bit.

Or imagine that your Great-Aunt Gertrude and your Great-Aunt Bertha are trying to work together on Thanksgiving dinner. Great-Aunt Gertrude is a proper Southern lady. She thinks no one should curse in mixed company (in fact, she’s rather suspicious of the word ‘goshdarnit’). She believes it is unconscionably rude for children not to say “sir” and “ma’am” to their elders.

Conversely, Great-Aunt Bertha skipped school in the fifties to go get drunk with sailors and was the first woman in the Hell’s Angels. Great-Aunt Bertha thinks it is very rude that Great-Aunt Gertrude keeps saying “a-HEM” five times a sentence just because she’s talking the way she normally talks. It’s not polite to interrupt what people are saying by getting offended and storming out. And that whole “sir” and “ma’am” business is *actually* offensive. Children are people and it is wrong to treat them as if they are subservient to adults.

Great-Aunt Bertha and Great-Aunt Gertrude will have some difficulty agreeing about what is polite behavior at the Thanksgiving table.

One could resolve these problems by taking some authority on etiquette, perhaps Miss Manners, and then saying that civility is officially now defined as doing what Miss Manners says to do. On the other hand, many aspects of etiquette have nothing to do with being nice to people but instead are ways of signalling that one is upper-class, or at least a middle-class person with pretensions of same. (Most obviously, anything about what forks one uses; more controversially, rules about greetings, introductions, when to bring gifts, etc.) You wind up excluding poor and less educated people, which people in many spaces don’t want.

So what’s the solution? There isn’t one that works literally 100% of the time. If you just give up on socially enforcing civility at all, then you get 4Chan. Not to bash 4Chan, but I for one am pretty happy about the existence of social spaces that are not 4Chan.

I think it’s important to think carefully about what your space is and is not for. Maybe this is actually just Great-Aunt Bertha’s Thanksgiving, and Great-Aunt Gertrude will have to suck up the curse words or organize her own Thanksgiving.

Sometimes you do want civil dialogue to occur between two groups who disagree a lot about what civility is. If everyone involved has good faith and is willing to compromise, that can happen okay. For example, maybe Great-Aunt Gertrude really cares about not hearing sex jokes, and Great-Aunt Bertha really cares about being allowed to swear, and they can have Thanksgiving together both feeling only a little bit uncomfortable.

If the rules are explicit (for example, in an online group with moderators), it’s a good idea to make sure all sides are equally represented in the group of people who enforce the rules, so everyone has their concerns respected. If the rules are implicit (for example, in a group of friends), it’s a good idea to focus mostly on correcting the behavior of people you agree with and *not* the behavior of people you disagree with. If you ever feel scared or defensive, take a break from the conversation: online, this

might mean stepping away from your computer, while offline you might ask for a change of subject.

Against Modest Epistemology

Follow-up to: Blind Empiricism

Modest epistemology doesn't need to reflect a skepticism about causal models as such. It can manifest instead as a wariness about putting weight down on *one's own* causal models, as opposed to others'.

In 1976, Robert Aumann demonstrated that two ideal Bayesian reasoners with the same priors cannot have common knowledge of a disagreement. Tyler Cowen and Robin Hanson have extended this result, establishing that even under various weaker assumptions, something has to go *wrong* in order for two agents with the same priors to get stuck in a disagreement.[1] If you and a trusted peer don't converge on identical beliefs once you have a full understanding of one another's positions, at least one of you must be making *some* kind of mistake.

If we were fully rational (and fully honest), then we would always eventually reach consensus on questions of fact. To become more rational, then, shouldn't we set aside our claims to special knowledge or insight and modestly profess that, really, we're all in the same boat?

When I'm trying to sort out questions like these, I often find it useful to start with a related question: "If I were building a brain from scratch, would I have it act this way?"

If I were building a brain and I expected it to have some non-fatal flaws in its cognitive algorithms, I expect that I would have it spend some of its time using those flawed reasoning algorithms to think about the world; and I would have it spend some of its time using those same flawed reasoning algorithms to better understand its reasoning algorithms. I would have the brain spend most of its time on object-level problems, while spending some time trying to build better meta-level models of its own cognition and how its cognition relates to its apparent success or failure on object-level problems.

If the thinker is dealing with a foreign cognitive system, I would want the thinker to try to model the other agent's thinking and *predict* the degree of accuracy this system will have. However, the thinker should also record the *empirical* outcomes, and notice if the other agent's accuracy is more or less than expected. If particular agents are more often correct than its model predicts, the system should recalibrate its estimates so that it won't be predictably mistaken in a known direction.

In other words, I would want the brain to reason about brains in pretty much the same way it reasons about other things in the world. And in practice, I suspect that the way I think, and the way I'd advise people in the real world to think, works very much like that:

- Try to spend most of your time thinking about the object level. If you're spending more of your time thinking about your own reasoning ability and competence than you spend thinking about Japan's interest rates and NGDP, or competing omega-6 vs. omega-3 metabolic pathways, you're taking your eye off the ball.

- Less than a majority of the time: Think about how reliable authorities seem to be and should be expected to be, and how reliable you are—using your own brain to think about the reliability and failure modes of brains, since that's what you've got. Try to be evenhanded in how you evaluate your own brain's *specific* failures versus the *specific* failures of other brains.[2] While doing this, *take your own meta-reasoning at face value*.

- … and then next, theoretically, should come the meta-meta level, considered yet more rarely. But I don't think it's necessary to develop *special* skills for meta-meta reasoning. You just apply the skills you already learned on the meta level to correct your own brain, and go on applying them *while* you happen to be meta-reasoning about who should be trusted, about degrees of reliability, and so on. Anything you've already learned about reasoning should automatically be applied to how you reason about meta-reasoning.[3]

- Consider whether someone else might be a better meta-reasoner than you, and hence that it might *not* be wise to take your own meta-reasoning at face value when disagreeing with them, *if you have been given strong local evidence to this effect*.

That probably sounded terribly abstract, but in practice it means that everything plays out in what I'd consider to be the obvious intuitive fashion.

i.

Once upon a time, my colleague Anna Salamon and I had a disagreement. I thought—this sounds really stupid in retrospect, but keep in mind that this was without benefit of hindsight—I thought that the best way to teach people about detaching from sunk costs was to write a script for local *Less Wrong* meetup leaders to carry out exercises, thus enabling all such meetups to be taught how to avoid sunk costs. We spent a couple of months trying to write this sunk costs unit, though a lot of that was (as I conceived of it) an up-front cost to figure out the basics of how a unit should work at all.

Anna was against this. Anna thought we should not try to carefully write a unit. Anna thought we should just find some volunteers and improvise a sunk costs teaching session and see what happened.

I explained that I wasn't starting out with the hypothesis that you *could* successfully teach anti-sunk-cost reasoning by improvisation, and therefore I didn't think I'd learn much from observing the improvised version fail. This may sound less stupid if you consider that I was accustomed to writing many things, most of which never worked or accomplished anything, and a very few of which people paid attention to and mentioned later, and that it had taken me years of writing practice to get even that far. And so, to me, negative examples seemed too common to be valuable. The literature was full of failed attempts to correct for cognitive biases—would one more example of that really help?

I tried to carefully craft a sunk costs unit that would rise above the standard level (which was failure), so that we would actually learn something when we ran it (I reasoned). I also didn't think up-front that it would be two months to craft; the

completion time just kept extending gradually—beware the planning fallacy!—and then at some point we figured we had to run what we had.

As read by one of the more experienced meetup leaders, the script did not work. It was, by my standards, a miserable failure.

Here are three lessons I learned from that experiment.

The first lesson is to not carefully craft anything that it was possible to *literally* just improvise and test immediately in its improvised version, ever. Even if the minimum improvisable product won't be representative of the real version. Even if you already expect the current version to fail. You *don't know* what you'll learn from trying the improvised version.[4]

The second lesson was that my model of teaching rationality by producing units for consumption at meetups wasn't going to work, and we'd need to go with Anna's approach of training teachers who could fail on more rapid cycles, and running centralized workshops using those teachers.

The third thing I learned was to avoid disagreeing with Anna Salamon in cases where we would have common knowledge of the disagreement.

What I learned wasn't quite as simple as, "Anna is often right." Eliezer is also often right.

What I learned wasn't as simple as, "When Anna and Eliezer disagree, Anna is more likely to be right." We've had a lot of first-order disagreements and I haven't particularly been tracking whose first-order guesses are right more often.

But the case above wasn't a first-order disagreement. I had presented my reasons, and Anna had understood and internalized them and given her advice, and *then* I had guessed that in a situation like this I was more likely to be right. So what I learned is, "Anna is sometimes right *even when my usual meta-reasoning heuristics say otherwise*," which was the real surprise and the first point at which something like an extra push toward agreement is additionally necessary.

It doesn't particularly surprise me if a physicist knows more about photons than I do; that's a case in which my usual meta-reasoning already predicts the physicist will do better, and I don't need any additional nudge to correct it. What I learned from that significant multi-month example was that my *meta-rationality*—my ability to judge which of two people is thinking more clearly and better integrating the evidence in a given context—was not particularly better than Anna's meta-rationality. And that meant the conditions for something like Cowen and Hanson's extension of Aumann's agreement theorem were actually being fulfilled. Not pretend ought-to-be fulfilled, but actually fulfilled.

Could adopting modest epistemology in general have helped me get the right answer in this case? The versions of modest epistemology I hear about usually involve deference to the majority view, to the academic mainstream, or to publicly recognized elite opinion. Anna wasn't a majority; there were two of us, and nobody else in particular was party to the argument. Neither of us were part of a mainstream. And at the point in time where Anna and I had that disagreement, any outsider would have thought that Eliezer Yudkowsky had the more impressive track record at teaching rationality. Anna wasn't yet heading CFAR. Any advice to follow track records, to trust externally observable eliteness in order to avoid the temptation to overconfidence,

would have favored listening to Yudkowsky over Salamon—that's part of the reason I trusted myself over her in the first place! And then I was wrong anyway, because in real life that is allowed to happen even when one person has more externally observable status than another.

Whereupon I began to hesitate to disagree with Anna, and hesitate even more if she had heard out my reasons and yet still disagreed with me.

I extend a similar courtesy to Nick Bostrom, who recognized the importance of AI alignment three years before I did (as I discovered afterwards, reading through one of his papers). Once upon a time I thought Nick Bostrom couldn't possibly get anything done in academia, and that he was staying in academia for bad reasons. After I saw Nick Bostrom successfully found his own research institute doing interesting things, I concluded that I was wrong to think Bostrom should leave academia—and also *meta-wrong* to have been so confident while disagreeing with Nick Bostrom. I still think that oracle AI (limiting AI systems to only answer questions) isn't a particularly useful concept to study in AI alignment, but every now and then I dust off the idea and check to see how much sense oracles currently make to me, because Nick Bostrom thinks they might be important even after knowing that I'm more skeptical.

There are people who think we all ought to behave this way toward each other as a matter of course. They reason:

a) on average, we can't all be more meta-rational than average; and

b) you can't trust the reasoning you use to think you're more meta-rational than average. After all, due to Dunning-Kruger, a young-Earth creationist will also think they have plausible reasoning for why they're more meta-rational than average.

… Whereas it seems to me that if I lived in a world where the average person on the street corner were Anna Salamon or Nick Bostrom, the world would look extremely different from how it actually does.

… And from the fact that you're reading this at all, I expect that if the average person on the street corner were *you*, the world would again look extremely different from how it actually does.

(In the event that this book is ever read by more than 30% of Earth's population, I withdraw the above claim.)

ii.

I once poked at someone who seemed to be arguing for a view in line with modest epistemology, nagging them to try to formalize their epistemology. They suggested that we all treat ourselves as having a black box receiver (our brain) which produces a signal (opinions), and treat other people as having other black boxes producing other signals. And we all received our black boxes at random—from an anthropic perspective of some kind, where we think we have an equal chance of being any observer. So we can't start out by believing that our signal is likely to be more accurate than average.

But I don't think of myself as having started out with the *a priori* assumption that I have a better black box. I learned about processes for producing good judgments, like Bayes's Rule, and this let me observe when other people violated Bayes's Rule, and try to keep to it myself. Or I read about sunk cost effects, and developed techniques for avoiding sunk costs so I can abandon bad beliefs faster. After having made observations about people's real-world performance and invested a lot of time and effort into getting better, I expect some degree of outperformance relative to people who haven't made similar investments.

To which the modest reply is: "Oh, but any crackpot could say that their personal epistemology is better because it's based on a bunch of stuff that they think is cool. What makes you different?"

Or as someone advocating what I took to be modesty recently said to me, after I explained why I thought it was sometimes okay to give yourself the discretion to disagree with mainstream expertise when the mainstream seems to be screwing up, in exactly the following words: "But then what do you say to the Republican?"

Or as Ozy Brennan puts it, in dialogue form:

> BECOMING SANE SIDE: "Hey! Guys! I found out how to take over the world using only the power of my mind and a toothpick."
>
> HARM REDUCTION SIDE: "You can't do that. Nobody's done that before."
>
> BECOMING SANE SIDE: "Of course they didn't, they were completely irrational."
>
> HARM REDUCTION SIDE: "But they thought they were rational, too."
>
> BECOMING SANE SIDE: "The difference is that I'm right."
>
> HARM REDUCTION SIDE: "They thought that, too!"

This question, "But what if a crackpot said the same thing?", I've never heard formalized—though it seems clearly central to the modest paradigm.

My first and primary reply is that there is a saying among programmers: "There is not now, nor has there ever been, nor will there ever be, any programming language in which it is the least bit difficult to write bad code."

This is known as Flon's Law.

The lesson of Flon's Law is that there is no point in trying to invent a programming language which can coerce programmers into writing code you approve of, because that is impossible.

The deeper message of Flon's Law is that this kind of defensive, adversarial, lock-down-all-the-doors, block-the-idiots-at-all-costs thinking doesn't lead to the invention of good programming languages. And I would say much the same about epistemology for humans.

Probability theory and decision theory shouldn't deliver clearly wrong answers. Machine-specified epistemology shouldn't mislead an AI reasoner. But if we're just dealing with verbal injunctions for humans, where there are degrees of freedom, then there is nothing we can say that a hypothetical crackpot could not somehow misuse.

Trying to defend against that hypothetical crackpot will not lead us to devise a good system of thought.

But again, let's talk formal epistemology.

So far as probability theory goes, a good Bayesian ought to condition on all of the available evidence. E. T. Jaynes lists this as a major desideratum of good epistemology—that if we know *A*, *B*, and *C*, we ought not to decide to condition only on *A* and *C* because we don't like where *B* is pointing. If you're trying to estimate the accuracy of your epistemology, and you know what Bayes's Rule is, then—on naive, straightforward, traditional Bayesian epistemology—you ought to condition on both of these facts, and estimate *P*(accuracy|know_Bayes) instead of *P*(accuracy). Doing anything other than that opens the door to a host of paradoxes.

The convergence that perfect Bayesians exhibit on factual questions doesn't involve anyone straying, even for a moment, from their individual best estimate of the truth. The idea isn't that good Bayesians try to make their beliefs more closely resemble their political rivals' so that their rivals will reciprocate, and it isn't that they toss out information about their own rationality. Aumann agreement happens *incidentally*, without any deliberate push toward consensus, through each individual's single-minded attempt to reason from their own priors to the hypotheses that best match their own observations (which happen to include observations about other perfect Bayesian reasoners' beliefs).

Modest epistemology seems to me to be taking the experiments on the outside view showing that typical holiday shoppers are better off focusing on their past track record than trying to model the future in detail, and combining that with the Dunning-Kruger effect, to argue that we ought to throw away most of the details in our self-observation. At its epistemological core, modesty says that we should abstract up to a particular *very general* self-observation, condition on it, and then not condition on anything else because that would be inside-viewing. An observation like, "I'm familiar with the cognitive science literature discussing which debiasing techniques work well in practice, I've spent time on calibration and visualization exercises to address biases like base rate neglect, and my experience suggests that they've helped," is to be generalized up to, "I use an epistemology which I think is good." I am then to ask myself what average performance I would expect from an agent, conditioning only on the fact that the agent is using an epistemology that they think is good, and not conditioning on that agent using Bayesian epistemology or debiasing techniques or experimental protocol or mathematical reasoning or anything in particular.

Only in this way can we force Republicans to agree with us… or something. (Even though, of course, anyone who wants to shoot off their own foot will actually just reject the whole modest framework, so we're not *actually* helping anyone who wants to go astray.)

Whereupon I want to shrug my hands helplessly and say, "But given that this isn't normative probability theory and I haven't seen modesty advocates appear to get any particular outperformance out of their modesty, why go there?"

I think that's my true rejection, in the following sense: If I saw a sensible formal epistemology underlying modesty and I saw people who advocated modesty going on to outperform myself and others, accomplishing great deeds through the strength of their diffidence, then, indeed, I would start paying very serious attention to modesty.

That said, let me go on beyond my true rejection and try to construct something of a *reductio*. Two *reductios*, actually.

The first *reductio* is just, as I asked the person who proposed the signal-receiver epistemology: “Okay, so why don’t you believe in God like a majority of people’s signal receivers tell them to do?”

“No,” he replied. “Just no.”

“What?” I said. “You’re allowed to say ‘just no’? Why can’t I say ‘just no’ about collapse interpretations of quantum mechanics, then?”

This is a serious question for modest epistemology! It seems to me that on the signal-receiver interpretation you have to believe in God. Yes, different people believe in different Gods, and you could claim that there’s a majority disbelief in every particular God. But then you could as easily disbelieve in quantum mechanics because (you claim) there isn’t a majority of physicists that backs any particular interpretation. You could disbelieve in the whole edifice of modern physics because no exactly specified version of that physics is agreed on by a majority of physicists, or for that matter, by a majority of people on Earth. If the signal-receiver argument doesn’t imply that we ought to average our beliefs together with the theists and all arrive at an 80% probability that God exists, or whatever the planetary average is, then I have no idea how the epistemological mechanics are supposed to work. If you’re allowed to say “just no” to God, then there’s clearly some level—object level, meta level, meta-meta level—where you are licensed to take your own reasoning at face value, despite a majority of other receivers getting a different signal.

But if we say “just no” to anything, even God, then we’re no longer modest. We are faced with the nightmare scenario of having *granted ourselves discretion* about when to disagree with other people, a discretionary process where we *take our own reasoning at face value*. (Even if a majority of others disagree about this being a good time to take our own beliefs at face value, telling us that reasoning about the incredibly deep questions of religion is surely the worst of all times to trust ourselves and our pride.) And then what do you say to the Republican?

And if you give people the license to decide that they ought to defer, e.g., only to a majority of members of the National Academy of Sciences, who mostly don’t believe in God; then surely the analogous license is for theists to defer to the true experts on the subject, their favorite priesthood.

The second *reductio* is to ask yourself whether a superintelligent AI system ought to soberly condition on the fact that, in the world so far, many agents (humans in psychiatric wards) have believed themselves to be much more intelligent than a human, and they have all been wrong.

Sure, the superintelligence thinks that it remembers a uniquely detailed history of having been built by software engineers and raised on training data. But if you ask any other random agent that thinks it’s a superintelligence, that agent will just tell you that it remembers a unique history of being chosen by God. Each other agent that believes itself to be a superintelligence will forcefully reject any analogy to the other humans in psychiatric hospitals, so clearly “I forcefully reject an analogy with agents who wrongly believe themselves to be superintelligences” is not sufficient justification to conclude that one really is a superintelligence. Perhaps the superintelligence will plead that its internal experiences, despite the extremely abstract and high-level point of similarity, are really extremely dissimilar in the details from those of the patient in

the psychiatric hospital. But of course, if you ask them, the psychiatric patient could just say the same thing, right?

I mean, the psychiatric patient *wouldn't* say that, the same way that a crackpot wouldn't *actually* give a long explanation of why they're allowed to use the inside view. But they *could*, and according to modesty, That's Terrible.

iii.

To generalize, suppose we take the following rule seriously as epistemology, terming it Rule M for Modesty:

> **Rule M**: Let X be a very high-level generalization of a belief subsuming specific beliefs X_1, X_2, X_3.... For example, X could be "I have an above-average epistemology," X_1 could be "I have faith in the Bible, and that's the best epistemology," X_2 could be "I have faith in the words of Mohammed, and that's the best epistemology," and X_3 could be "I believe in Bayes's Rule, because of the Dutch Book argument." Suppose that all people who believe in any X_i, taken as an entire class X, have an average level F of fallibility. Suppose also that most people who believe some X_i also believe that their X_i is not similar to the rest of X, and that they are not like most other people who believe some X, and that they are less fallible than the average in X. Then when you are assessing your own expected level of fallibility you should condition only on being in X, and compute your expected fallibility as F. You should not attempt to condition on being in X_3 or ask yourself about the average fallibility you expect from people in X_3.

Then the first machine superintelligence should conclude that it is in fact a patient in a psychiatric hospital. And you should believe, with a probability of around 33%, that you are currently asleep.

Many people, while dreaming, are not aware that they are dreaming. Many people, while dreaming, may believe at some point that they have woken up, while still being asleep. *Clearly* there can be no license from "I think I'm awake" to the conclusion that you actually are awake, since a dreaming person could just dream the same thing.

Let Y be the state of not thinking that you are dreaming. Then Y_1 is the state of a dreaming person who thinks this, and Y_2 is the state of actually being awake. It boots nothing, on Rule M, to say that Y_2 is introspectively distinguishable from Y_1 or that the inner experiences of people in Y_2 are actually quite different from those of people in Y_1. Since people in Y_1 usually falsely believe that they're in Y_2, you ought to just condition on being in Y, not condition on being in Y_2. Therefore you should assign a 67% probability to currently being awake, since 67% of observer-moments who believe they're awake are actually awake.

Which is why—in the distant past, when I was arguing against the modesty position for the first time—I said: "Those who dream do not know they dream, but when you are awake, you know you are awake." The modest haven't formalized their epistemology very much, so it would take me some years past this point to write down the Rule M that I thought was at the heart of the modesty argument, and say that "But you know you're awake" was meant to be a *reductio* of Rule M in particular, and why.

Reasoning under uncertainty and in a biased and error-prone way, still we can say that the probability we're awake isn't just a function of how many awake versus sleeping people there are in the world; and the rules of reasoning that let us update on Bayesian evidence that we're awake can serve *that* purpose equally well whether or not dreamers can profit from using the same rules. If a rock wouldn't be able to use Bayesian inference to learn that it is a rock, still I can use Bayesian inference to learn that I'm not.

Next: **Status Regulation and Anxious Underconfidence**.

---

1. See Cowen and Hanson, "Are Disagreements Honest?" ↩
2. This doesn't mean the net estimate of who's wrong comes out 50-50. It means that if you rationalized last Tuesday then you expect yourself to rationalize this Tuesday, if you would expect the same thing of someone else after seeing the same evidence. ↩
3. And then the recursion stops here, first because we already went in a loop, and second because in practice nothing novel happens after the third level of any infinite recursion. ↩
4. Chapter 22 of my *Harry Potter* fanfiction, *Harry Potter and the Methods of Rationality*, was written after I learned this lesson. ↩

Qualitative differences

Probably I am making a huge mistake resuming a ten years old discussion, especially because this is my first post. Anyway, let's make this experiment.

One common argument for choosing "dust specks" over "torture" is that the experience of torture is qualitatively different form the experience of receiving a dust speck in the eye and so the two should not be compared (assuming that there won't be other long term negative consequences such as car accidents and surgical mistakes).

A moderate pain doesn't cause the same reactions in the human organism that happen during an extreme and prolonged distress such as: panic attacks, alterations of physiological functions and psychological trauma with all the capacity impairment that derive from it.

But the most obvious characteristic that distinguishes the pain caused by a torture from the pain caused by a dust speck in the eye is the intolerability: people would rather die than feel 50 years of torture. Some argue that these qualitative differences don't exist because we can create a sequence of injures each comparable to the previous and the next.

For example:

30 dust specks divided among 30 people≃1 slap.

10 slaps divided among 10 people≃1 punch.

10 punches divided among 10 people≃1 small cut.

10 small cuts divided among 10 people≃1 deep cut.

10 deep cuts divided among 10 people≃1 tonsillectomy without anesthesia.

10 tonsillectomies without anesthesia divided among 10 people≃1 torture.

So it seems that we can't clearly distinguish what is tolerable from what is intolerable, unless we choose on the spectrum of the grades of pains an arbitrary limit, or rather a distance, that separate two incomparable kinds of injuries. Many have rejected this idea because the line would be subjective and because the difference between things near the limit is small.

However these limits have an important function in our decision making process: in some situations there could be a range of options which are similar among each other, but at same time have differences which accumulate step by step, until they become too relevant to be ignored.

Consider this mind experiment: a man has bought a huge house with many rooms, is favorite color is orange and he wants to paint all the rooms with it. Omega offers its help and gives five options, it will paint:

A) One room with R=255 G=150 B=0 paint

B) Two rooms with R=255 G=120 B=0 paint

C) Four rooms with R=255 G=90 B=0 paint

D) Eight rooms with R=255 G=60 B=0 paint

E) Sixteen rooms with R=255 G=30 B=0 paint.

What will the man prefer? Probably not the first option, but also not the last. He will find a balance between his desire to have orange rooms and his laziness. But what would the man choose if the only possible options were A) and E)?

Probably the first one: although red and orange are part of a continuous spectrum and he doesn't know exactly where their separation line is, he can still distingue the two color, and no amount of red is comparable to his favorite color.

Also the lifespan dilemma shows that for many people the answer can't be just a matter of expected value, otherwise everyone would agree on reducing the probability of success to values near to 0. When making a choice for the dilemma, people consider two emotions: the desire to live longer and the fear to die, the answer is found when people think that they have reached the line which separates an acceptable risk from an excessive one.

This line is quite arbitrary and subjective, like the separation line between red and orange. Nevertheless, most people can see that the colors of ripe strawberries and ripe oranges are different, that a minimum amount of safety is required and that annoyance and agony are different things, even if there are intermediate situations between them.

One last example: the atmosphere is divided in many layers but the limits of each layer are nebulous because their characteristics get more similar near their edges; moreover the last layer is made of very low density gas whose particles constantly escape into space. Nevertheless if we choose two points in the atmosphere we can see that their characteristics became more different as we increase the difference of their altitude, until we can affirm that they are in two different layer.

This distance is difficult to define and people will disagree on its precise length, however we cannot ignore the difference between things that are very distant from each other, as we cannot ignore that atmosphere is still present at 1km of altitude while it is not present at 100,000km of altitude, despite the fact that its outer limit is very ill defined.

Similarly we can use the knowledge of human physiology during the events of life to choose a distance that separate two incomparable kinds of experiences, and so give the priority to the actions that can really change people life quality. For instance I think it is preferable to give 1,000,000$ to a poor family rather than 3^^^3$ to 3^^^3 middle class families, because each single dollar in the latter case would cause an irrelevant change compared to the former case. In other words only the money in the first case will extinguish hunger.

Personally my empathy usually makes me prefer utilitarian, or rather quantitative choices, when the qualities of experiences are similar, but it also compels me to save one person from something intolerable, rather than 3^^^3 people from something tolerable. This is not scope insensitivity, but rather coherence: I could stand 3^^^3 annoyed but bearable lives, but not 3^^^3-1 not annoyed lives plus one unbearable life, which is quite tautological when you consider the definition of "unbearable".

I don’t know exactly where my cut-off is, and I won’t be outraged if someone else’s cut-off is higher or lower, but I would be rather skeptical if someone claimed that he can’t bear a dust specks (or rather 3^^^3 dust speck diluted among 3^^^3 lives).

In addition, if we consider our utility to be something more than raw pleasure, then our disutility should be more than raw pain. Surely pleasure is valuable, but many people want something more from life, and for them no amount of pleasure can substitute the value of discovery, accomplishment, relationships, experiences, creativity, etc. All this things are valuable but only together they make something qualitatively more important, which is often called human flourishing or Eudaimonia.

Similarly only if pain is combined with frustration, fear, desperation, panic, etc. it becomes something qualitatively worst such as agony. Eudaimonia can resist a dust speck, or even a stubbed toe, especially since we have the ability to heal, find relief, and even become stronger after a tolerable stress, but it can’t resist 50 years of torture.

The distinction between pain and agony is the same between colours or between pain and pleasure: in each case neurons fires but in different ways, and pain is not the mere inhibition of pleasure centre, after all one can be happy also when feeling a moderate pain.

Utilitarianism has two orthogonal goals: minimize pain and maximize pleasure. People that choose dust specks could simply have another one: minimize agony. If physiological reactions can be used to distinguish pain from pleasure, then why shouldn’t we consider other characteristics and be more precise when considering people values?

Recognizing the importance of qualitative differences and physiological limits surely makes the utility calculus more complicate. However there are also advantages: the result would be a system more compatible with people intuitions, more egalitarian, and more safe from utility monsters.

One final thought: I don’t delude myself that I will change someone opinion on the problem. After all, people who choose torture consider humans intuitions wrong since they go against utilitarianism.

One the other hand people who choose dust specks consider utilitarianism (or at least some forms of utilitarianism) wrong because it goes against moral intuitions, and because we think that to simplify human morality is dangerous.

People in the first case care about maximizing pleasure, and minimizing pain.

We “dust specks chooser” have another aim. This aim is probably to create an ethical system that that establishes a safety net to protect people from unbearable and situations and from injuries that really destroy the capability to pursue a decent, serene, worthy life, which probably is the foremost goal of most people.

For utilitarianism is morally right to torture billions of people for 50 years if this will give someone 3^^^^^3 years of happiness. There is no reason why we should endorse something so repugnant for the sake of simplicity: utilitarianism is an oversimplification, it deliberately ignores equality and justice which are part of people morality as much as compassion.

I don’t deny that utilitarianism usually works, however to shut up and multiply is an act of blind faith: it makes people stick to the unjust parts of system rather than reform it, and I can’t deny that I am rather worried, but also curios about the origins of this zeal.

I would really like to be able to discuss with some torture chooser. I am sure that it could be an interesting mutual occasion to learn more about minds that are very different from our own.

P.S. I am not from an English speaking country. The assertive manner in which I wrote this post was merely a way to expose my ideas more simply, any advice about style or grammar is appreciated.

The Darwin Game

Epistemic Status: True story

The plan is that this post will begin the sequence Zbybpu'f Nezl.

In college I once took a class called Rational Choice. Because obviously.

Each week we got the rules for, played and discussed a game. It was awesome.

For the grand finale, and to determine the winner of the prestigious Golden Shark Award for best overall performance, we submitted computer program architectures (we'd tell the professor what our program did, within reason, and he'd code it for us) to play The Darwin Game.

The Darwin Game is a variation slash extension of the iterated prisoner's dilemma. It works like this:

For the first round, each player gets 100 copies of their program in the pool, and the pool pairs those programs at random. You can and often will play against yourself.

Each pair now plays an iterated prisoner's dilemma variation, as follows. Each turn, each player simultaneously submits a number from 0 to 5. If the two numbers add up to 5 or less, both players earn points equal to their number. If the two numbers add up to 6 or more, neither player gets points. This game then lasts for a large but unknown number of turns, so no one knows when the game is about to end; for us this turned out to be 102 turns.

Each pairing is independent of every other pairing. You do not know what round of the game it is, whether you are facing a copy of yourself, or any history of the game to this point. Your decision algorithm does the same thing each pairing.

At the end of the round, all of the points scored by all of your copies are combined. Your percentage of all the points scored by all programs becomes the percentage of the pool your program gets in the next round. So if you score 10% more points, you get 10% more copies next round, and over time successful programs will displace less successful programs. Hence the name, The Darwin Game.

Your goal is to have as many copies in the pool at the end of the 200th round as possible, or failing that, to survive as many rounds as possible with at least one copy.

If both players coordinate to split the pot, they will score 2.5 per round.

To create some common terminology for discussions, 'attack' or means to submit 3 (or a higher number) more than half the time against an opponent willing to not do that, and to 'fold' or 'surrender' is to submit 2 (or a lower number) more than half the time, with 'full surrender' being to always submit 2. To 'cooperate' is to alternate 2 and 3 such that you each score 2.5 per round.

In this particular example we expected and got about 30 entries, and I was in second place in the points standings, so to win The Golden Shark, I had to beat David by a substantial amount and not lose horribly to the students in third or fourth.

What program do you submit?

(I recommend actually taking some time to think about this before you proceed.)

Some basic considerations I thought about:

1. The late game can come down to very small advantages that compound over time.

2. You need to survive the early game *and* win the late game. This means you need to succeed in a pool of mostly-not-smart programs, and then win in a pool of smart programs, and then in a pool of smart programs that outlasted other smart programs.

3. Scoring the maximum now regardless of what your opponent scores helps you early, but kills you late. In the late game, not letting your opponent score more points than you is very important, especially once you are down to two or three programs.

4. In the late game, how efficiently you cooperate with yourself is very important.

5. Your reaction to different programs in the mid game will help determine your opponents in the end game. If an opponent that outscores you in a pairing survives into the late game, and co-operates with itself, you lose.

6. It is all right to surrender, even fully surrender, to an opponent if and only if they will be wiped out by other programs before you become too big a portion of the pool, provided you can’t do better.

7. It is much more important to get to a good steady state than to get there quickly, although see point one. Getting people to surrender to you would be big game.

8. Some of the people entering care *way* more than others. Some programs will be complex and consider many cases and be trying hard to win, others will be very simple and not trying to be optimal.

9. It is hard to tell what others will interpret as cooperation and defection, and it might be easy to accidentally make them think you’re attacking them.

10. There will be some deeply silly programs out there at the start. One cannot assume early programs are doing remotely sensible things.

That leaves out many other considerations, including at least one central one. Next time, I’ll go over what happened on game day.

Blind Empiricism

Follow-up to: Living in an Inadequate World

The thesis that needs to be contrasted with modesty is not the assertion that everyone can beat their civilization all the time. It's not that we should be *the sort of person* who sees the world as mad and pursues the strategy of believing a hot stock tip and investing everything.

It's just that it's *okay* to reason about the particulars of where civilization might be inadequate, *okay* to end up believing that you can state a better monetary policy than the Bank of Japan is implementing, *okay* to check that against observation whenever you get the chance, and okay to update on the results in *either* direction. It's okay to act on a model of what you think the rest of the world is good at, and for this model to be sensitive to the specifics of different cases.

Why might this *not* be okay?

It could be that "acting on a model" is suspect, at least when it comes to complicated macrophenomena. Consider Isaiah Berlin's distinction between "hedgehogs" (who rely more on theories, models, global beliefs) and "foxes" (who rely more on data, observations, local beliefs). Many people I know see the fox's mindset as more admirable than the hedgehog's, on the basis that it has greater immunity to fantasy and dogmatism. And Philip Tetlock's research has shown that political experts who rely heavily on simple overarching theories—the kind of people who use the word "moreover" more often than "however"—perform substantially worse on average in forecasting tasks.[1]

Or perhaps the suspect part is when models are "sensitive to the specifics of different cases." In a 2002 study, Buehler, Griffin, and Ross asked a group of experimental subjects to provide lots of details about their Christmas shopping plans: where, when, and how. On average, this experimental group expected to finish shopping more than a week before Christmas. Another group was simply asked when they expected to finish their Christmas shopping, with an average response of 4 days. Both groups finished an average of 3 days before Christmas. Similarly, students who expected to finish their assignments 10 days before deadline actually finished one day before deadline; and when asked when they had previously completed similar tasks, replied, "one day before deadline." This suggests that taking the *outside view* is an effective response to the planning fallacy: rather than trying to predict how many hiccups and delays your plans will run into by reflecting in detail on each plan's particulars (the "inside view"), you can do better by just guessing that your future plans will work out roughly as well as your past plans.

As stated, these can be perfectly good debiasing measures. I worry, however, that many people end up misusing and overapplying the "outside view" concept very soon after they learn about it, and that a lot of people tie too much of their mental conception of what good reasoning looks like to the stereotype of the humble empiricist fox. I recently noticed this as a common thread running through three conversations I had.

I am not able to recount these conversations in a way that does justice to the people I spoke to, so please treat my recounting as an unfair and biased illustration of relevant ideas, rather than as a neutral recitation of the facts. My goal is to illustrate the kinds of reasoning patterns I think are causing epistemic harm: to point to some canaries in the coal mine, and to be clear that when I talk about modesty I'm not just talking about Hal Finney's majoritarianism or the explicit belief in civilizational adequacy.

i.

Conversation 1 was about the importance of writing code to test AI ideas. I suggested that when people tried writing code to test an idea I considered important, I wanted to see the code in advance of the experiment, or without being told the result, to see if I could predict the outcome correctly.

I got pushback against this, which surprised me; so I replied that my having a chance to make advance experimental predictions was important, for two reasons.

First, I thought it was important to develop a skill and methodology of predicting "these sorts of things" in advance, because past a certain level of development when working with smarter-than-human AI, if you can't see the bullets coming in advance of the experiment, the experiment kills you. This being the case, I needed to test this skill as much as possible, which meant trying to make experimental predictions in advance so I could put myself on trial.

Second, if I could predict the results correctly, it meant that the experiments weren't saying anything I hadn't figured out through past experience and theorizing. I was worried that somebody might take a result I considered an obvious prediction under my current views and say that it was evidence against my theory or methodology, since both often get misunderstood.[2] If you want to use experiment to show that a certain theory or methodology fails, you need to give advocates of the theory/methodology a chance to say beforehand what they think they predict, so the prediction is on the record and neither side can move the goalposts.

And I still got pushback, from a MIRI supporter with a strong technical background; so I conversed further.

I now suspect that—at least this is what I think was going on—their mental contrast between empiricism and theoreticism was so strong that they thought it was unsafe to have a theory *at all*. That having a theory made you a bad hedgehog with one big idea instead of a good fox who has lots of little observations. That the dichotomy was between making an advance prediction *instead of doing the experiment*, versus doing the experiment *without any advance prediction*. Like, I suspect that every time I talked about "making a prediction" they heard "making a prediction instead of doing an experiment" or "clinging to what you predict will happen and ignoring the experiment."

I can see how this kind of outlook would develop. The policy of making predictions to *test* your understanding, to put it on trial, presupposes that you can execute the "quickly say *oops* and abandon your old belief" technique, so that you can employ it if the prediction turns out to be wrong. To the extent that "quickly say *oops* and abandon your old belief" is something the vast majority of people fail at, maybe on an

individual level it's better for people to try to be pure foxes and only collect observations and try not to have any big theories. Maybe the average cognitive use case is that if you have a big theory and observation contradicts it, you will find some way to keep the big theory and thereby doom yourself. (The "Mistakes Were Made, But Not By Me" effect.)

But from my perspective, there's no choice. You just have to master "say *oops*" so that you can have theories and make experimental predictions. Even on a strictly empiricist level, if you aren't allowed to have models and you don't make your predictions in advance, you learn less. An empiricist of that sort can only learn surface generalizations about whether this phenomenon superficially "looks like" that phenomenon, rather than building causal models and putting them on trial.

ii.

Conversation 2 was about a web application under development, and it went something like this.

STARTUP FOUNDER 1: I want to get (primitive version of product) in front of users as fast as possible, to see whether they want to use it or not.

ELIEZER: I predict users will not want to use this version.

FOUNDER 1: Well, from the things I've read about startups, it's important to test as early as possible whether users like your product, and not to overengineer things.

ELIEZER: The concept of a "minimum viable product" isn't the minimum product that compiles. It's the least product that is the *best tool in the world* for some particular task or workflow. If you don't have an MVP in that sense, of course the users won't switch. So you don't have a testable hypothesis. So you're not really learning anything when the users don't want to use your product.[3]

FOUNDER 1: No battle plan survives contact with reality. The important thing is just to get the product in front of users as quickly as possible, so you can see what they think. That's why I'm disheartened that (group of users) did not want to use (early version of product).

ELIEZER: This reminds me of a conversation I had about AI twice in the last month. Two separate people were claiming that we would only learn things empirically by experimenting, and I said that in cases like that, I wanted to see the experiment description in advance so I could make advance predictions and put on trial my ability to foresee things without being hit over the head by them.

In both of those conversations I had a very hard time conveying the idea, "Just because I have a theory does not mean I have to be insensitive to evidence; the evidence *tests* the theory, potentially falsifies the theory, but for that to work you need to make experimental predictions in advance." I think I could have told you in advance that (group of users) would not want to use (early version of product), because (group of users) is trying to accomplish (task 1) and this version of the product is not the best available tool they'll have seen for doing (task 1).

I can’t convey it very well with all the details redacted, but the impression I got was that the message of “distrust theorizing” had become so strong that Founder 1 had stopped trying to model users in detail and thought it was futile to make an advance prediction. But if you can’t model users in detail, you can’t think in terms of workflows and tasks that users are trying to accomplish, or at what point you become visibly the best tool the user has ever encountered to accomplish some particular workflow (the minimum viable product). The alternative, from what I could see, was to think in terms of “features” and that as soon as possible you would show the product to the user and see if they wanted that subset of features.

There’s a version of this hypothesis which does make sense, which is that when you have the minimum compilable product that it is physically possible for a user to interact with, you can ask one of your friends to sit down in front of it, you can *make a prediction* about what parts they will dislike or find difficult, and then you can see if your prediction is correct. Maybe your product actually fails much earlier than you expect.

But this is not like getting early users to voluntarily adopt your product. This is about observing, as early as possible, how volunteers react to unviable versions of your product, so you know what needs fixing earliest or whether the exposed parts of your theory are holding up so far.

It really looks to me like the modest reactions to certain types of overconfidence or error are taken by many believers in modesty to mean, in practice, that theories just get you into trouble; that you can either make predictions *or* look at reality, but not both.

iii.

Conversation 3 was with Startup Founder 2, a member of the effective altruism community who was making Material Objects—I’ll call them “Snowshoes”—who had remarked that modern venture capital was only interested in 1000x returns and not 20x returns.

I asked why he wasn’t trying for 1000x returns with his current company selling Snowshoes—was that more annoyance/work than he wanted to undertake?

He replied that most companies in a related industry, Flippers, weren’t that large, and it seemed to him that based on the outside view, he shouldn’t expect his company to become larger than the average company in the Flippers industry. He asked if I was telling him to try being more confident.

I responded that, no, the thing I wanted him to think was orthogonal to modesty versus confidence. I observed that the customer use case for Flippers was actually quite different from Snowshoes, and asked him if he’d considered how many uses of Previous Snowshoes in the world would, in fact, benefit from being replaced by the more developed version of Snowshoes he was making.

He said that this seemed to him too much like optimism or fantasy, compared to asking what his company had to do next.

I had asked about how customers would benefit from new and improved Snowshoes because my background model says that startups are more likely to succeed if they provide real economic value—value of the kind that Danslist would provide over Craigslist if Danslist succeeded, and of the kind that Craigslist provides over newspaper classifieds. Getting people to actually buy your product, of course, is a separate question from whether it would provide real value of that kind. And there's an obvious failure mode where you're in love with your product and you overestimate the product's value or underestimate the costs to the user. There's an obvious failure mode where you just look at the real economic value and get all cheerful about that, without asking the further necessary question of how many decisionmakers *will* choose to use your product; or whether your marketing message is either opaque or easily faked; or whether any competitors will get there first if they see you being successful early on; or whether you could defend a price premium in the face of competition. But the question of real economic value seems to me to be one of the factors going into a startup's odds of succeeding—Craigslist's success is in part explained by the actual benefit buyers and sellers derive from the existence of Craigslist—and worth factoring out before discussing purchaser decisionmaking and value-capturing questions.[4]

It wasn't that I was trying to get Founder 2 to be more optimistic (though I did think, given his Snowshoes product, that he ought to at least *try* to be more ambitious). It was that it looked to me like the outside view was shutting down his causal model of how and why people might use his product, and substituting, "Just try to build your Snowshoes and see what happens, and at best don't expect to succeed more than the average company in a related industry." But I don't think you can get so far as even the *average* surviving company, unless you have a causal model (the dreaded inside view) of where your company is supposed to go and what resources are required to get there.

I was asking, "What level do you want to grow to? What needs to be done for your company to grow that much? What's the obstacle to taking the next step?" And… I think it felt immodest to him to claim that his company could grow to a given level; so he thought only in terms of things he knew he could try, forward-chaining from where he was rather than backward-chaining from where he wanted to go, because that way he didn't need to immodestly think about succeeding at a particular level, or endorse an inside view of a particular pathway.

I think the details of his business plan had the same outside-view problem. In the Flippers industry, two common versions of Flippers that were sold were Deluxe Flippers and Basic Flippers. Deluxe Flippers were basically preassembled Basic Flippers, and Deluxe Flippers sold for a much higher premium than Basic Flippers even though it was easy to assemble them.

We were talking about a potential variation of his Snowshoes, and he said that it would be too expensive to ship a Deluxe version, but not worth it to ship a Basic version, given the average premiums the outside view said these products could command.

I asked him *why*, in the Flippers industry, Deluxe sold for such a premium over Basic when it was so easy to assemble Basic into Deluxe. Why was this price premium being maintained?

He suggested that maybe people really valued the last little bit of convenience from buying Deluxe instead of Basic.

I suggested that in this large industry of slightly differentiated Flippers, maybe a lot of price-sensitive consumers bought only Basic versions, meaning that the few Deluxe buyers were price-insensitive. I then observed again that the best use case for his product was quite different from the standard use case in the Flipper industry, and that he didn't have much direct competition. I suggested that, for his customers that weren't otherwise customers in the Flippers industry, it wouldn't make much of a difference to his pricing power whether he sold Deluxe or the much easier to ship Basic version.

And I remarked that it seemed to me unwise in general to look at a *mysterious* pricing premium, and assume that you could get that premium. You couldn't just look at average Deluxe prices and assume you could get them. Generally speaking, this indicates some sort of rent or market barrier; and where there is a stream of rent, there will be walls built to exclude other people from drinking from the stream. Maybe the high Deluxe prices meant that Deluxe consumers were hard to market to, or very unlikely to switch providers. You couldn't just take the outside view of what Deluxe products tended to sell like.

He replied that he didn't think it was wise to say that you had to fully understand every part of the market before you could do anything; especially because, if you had to understand why Deluxe products sold at a premium, it would be so easy to just make up an explanation.

Again I understand where he was coming from, in terms of the average cognitive use case. When I try to explain a phenomenon, I'm also implicitly relying on my ability to use a technique like "don't even start to rationalize," which is a skill that I started practicing at age 15 and that took me a decade to hone to a reliable and productive form. I also used the "notice when you're confused about something" technique to ask the question, and a number of other mental habits and techniques for explaining mysterious phenomena—for starters, "detecting goodness of fit" (see whether the explanation feels "forced") and "try further critiquing the answer." Maybe there's no point in trying to explain why Deluxe products sell at a premium to Basic products, if you don't already have a lot of cognitive technique for not coming up with terrible explanations for mysteries, along with enough economics background to know which things are important mysteries in the first place, which explanations are plausible, and so on.

But at the same time, it seems to me that there is a learnable skill here, one that entrepreneurs and venture capitalists at least *have* to learn if they want to succeed on purpose instead of by luck.

One needs to be able to identify mysterious pricing and sales phenomena, read enough economics to speak the right simplicity language for one's hypotheses, and then not come up with terrible rationalizations. One needs to learn the key answers for how the challenged industry works, which means that one needs to have explicit hypotheses that one can test as early as possible.

Otherwise you're… not quite doomed *per se*, but from the perspective of somebody like me, there will be ten of you with bad ideas for every one of you that happens to have a good idea. And the people that do have good ideas will not really understand what human problems they are addressing, what their potential users' relevant motivations are, or what are their critical obstacles to success.

Given that analysis of ideas takes place on the level it does, I can understand why people would say that it's futile to try to analyze ideas, or that teams rather than ideas are important. I'm not saying that either entrepreneurs or venture capitalists could, by an effort of will, suddenly become great at analyzing ideas. But it seems to me that the outside view concept, along with the Fox=Good/Hedgehog=Bad, Observation=Good/Theory=Bad messages—including the related misunderstanding of MVP as "just build something and show it to users"—are preventing people from even starting to develop those skills. At least, my observation is that some people go too far in their skepticism of model-building.[5]

Maybe there's a valley of bad rationality here and the injunction to not try to have theories or causal models or preconceived predictions is protective against entering it. But first, if it came down to only those alternatives, I'd frankly rather see twenty aspiring rationalists fail painfully until one of them develops the required skills, rather than have nobody with those skills. And second, god damn it, there has to be a better way.

iv.

In situations that are drawn from a barrel of causally similar situations, where human optimism runs rampant and unforeseen troubles are common, the outside view beats the inside view. But in novel situations where causal mechanisms differ, the outside view fails—there may not be relevantly similar cases, or it may be ambiguous which similar-looking cases are the right ones to look at.

Where two sides disagree, this can lead to *reference class tennis*—both parties get stuck insisting that their own "outside view" is the correct one, based on diverging intuitions about what similarities are relevant. If it isn't clear what the set of "similar historical cases" is, or what conclusions we should draw from those cases, then we're forced to use an inside view—thinking about the causal process to distinguish relevant similarities from irrelevant ones.

You shouldn't avoid outside-view-style reasoning in cases where it looks likely to work, like when planning your Christmas shopping. But in many contexts, the outside view simply can't compete with a good theory.

Intellectual progress on the whole has usually been the process of moving from surface-level resemblances to more technical understandings of particulars. Extreme examples of this are common in science and engineering: the deep causal models of the world that allowed humans to plot the trajectory of the first moon rocket before launch, for example, or that allow us to verify that a computer chip will work before it's ever manufactured.

Where items in a reference class differ causally in more ways than two Christmas shopping trips you've planned or two university essays you've written, or where there's temptation to cherry-pick the reference class of things you consider "similar" to the phenomenon in question, or where the particular biases underlying the planning fallacy just aren't a factor, you're often better off doing the hard cognitive labor of building, testing, and acting on models of how phenomena actually work, even if those models are very rough and very uncertain, or admit of many exceptions and nuances. And, of course, during and after the construction of the model, you have

to look at the data. You still need fox-style attention to detail—and you certainly need empiricism.

The idea isn't, "Be a hedgehog, not a fox." The idea is rather: developing accurate beliefs requires both observation of the data *and* the development of models and theories that can be tested by the data. In most cases, there's no real alternative to sticking your neck out, even knowing that reality might surprise you and chop off your head.

Next: **Against Modest Epistemology**.

The full book will be available November 16th. You can go to equilibriabook.com to pre-order the book, or sign up for notifications about new chapters and other developments.

---

1. See Philip Tetlock, "Why Foxes Are Better Forecasters Than Hedgehogs." ↩

2. As an example, my conception of the reward hacking problem for reinforcement learning systems is that below certain capability thresholds, making the system smarter will often produce increasingly helpful behavior, assuming the rewards are a moderately good proxy for the actual objectives we want the system to achieve. The problem of the system exploiting loopholes and finding ways to maximize rewards in undesirable ways is mainly introduced when the system's resourcefulness is great enough, and its policy search space large enough, that operators can't foresee even in broad strokes what the reward-maximizing strategies are likely to look like. If this idea gets rounded off to just "making an RL system smarter will always reduce its alignment with the operator's goal," however, then a researcher will misconstrue what counts as evidence for or against prioritizing reward hacking research.

   And there are many other cases where ideas in AI alignment tend to be misunderstood, largely because "AI" calls to mind present-day applications. It's certainly possible to run useful experiments with present-day software to learn things about future AGI systems, but "see, this hill-climbing algorithm doesn't exhibit the behavior you predicted for highly capable Bayesian reasoners" will usually reflect a misconception about what the concept of Bayesian reasoning is doing in AGI alignment theory. ↩

3. I did not say this then, but I should have: Overengineering is when you try to make everything look pretty, or add additional cool features that you think the users will like… not when you try to put in the key core features that are necessary for your product to be the best tool the user has ever seen for at least one workflow. ↩

4. And a startup founder definitely needs to ask that question and answer it before they go out and try to raise venture capital from investors who are looking for 1000x returns. Don't discount your company's case before it starts. They'll do that for you. ↩

5. As Tetlock puts it in a discussion of the limitations of the fox/hedgehog model in the book *Superforecasting*: "Models are supposed to simplify things, which is why even the best are flawed. But they're necessary. Our minds are full of models. We couldn't function without them. And we often function pretty well because some of our models are decent approximations of reality." ↩

The Right to be Wrong

Epistemic Status: pretty confident

Zvi recently came out with a post "You Have the Right to Think", in response to Robin Hanson's "Why be Contrarian?", itself a response to Eliezer Yudkowsky's new book *Inadequate Equilibria.* All of these revolve around the question of when you should think you "know better" or can "do better" than the status quo of human knowledge or accomplishment. But I think there's a lot of conflation of different kinds of "should" going on.

Yudkowsky's book, and Hanson's post, are mostly about *epistemic* questions — when are you likely to get the right answer by examining an issue yourself vs. trusting experts?

Inadequate Equilibria starts with the canonical example of when you *can't* outperform the experts — betting on the stock market — and explains about efficient markets, and then goes on to look into what kinds of situations deviate from efficient markets such that an individual could outperform the collective intelligence of everyone who's tried so far. For instance, you might well be able to find a DIY treatment for your health problem that works better than anything your doctor would prescribe you, in certain situations — but due to the same incentive problems that prevented medical consensus from finding that treatment, you probably wouldn't be able to get it to become the standard of care in the mass market.

Hanson mostly agrees with Yudkowsky's analysis, except on some points where he thinks the argument for individual judgment being reliable is weaker.

Zvi seems to be talking about a different thing altogether when he talks about the "rights" that people have.

When he says "You have the right to disagree even when others would not, given your facts and reasoning, update their beliefs in your direction" or "You have the right to believe that someone else has superior meta-rationality and all your facts and reasoning, and still disagree with them", I assume he's not saying that you'd be *more likely to get the right answer to a question* in such cases — I think that would be false. If we posit someone who knows better than me in every relevant way, I'd definitionally be more likely to get the right answer by listening to her than disagreeing with her!

So, what does it mean to have a right to disagree *even when it makes you more likely to be wrong?* How can you have a right to be wrong?

I can think of two simple meanings and one subtle meaning.

The Right To Your Opinion

The first sense in which you "have a right to be wrong" is social and psychological.

It's a basic tenet of free and pluralistic societies that you have the *legal* right to believe a false thing, and express your belief. It is not a crime to write a horoscope column. You can't be punished by force just for being wrong. "Bad argument gets counterargument. Does not get bullet. Never. Never ever never for ever."

And tolerant, pluralist cultures generally don't believe in doing too much *social* punishment of people for being wrong, either. It's human to make mistakes; it's normal for people to disagree and not be able to resolve the disagreement; if you shame people as though being wrong is horribly taboo, your community is going to be a more disagreeable and stressful place. (Though some communities are willing to make that tradeoff in exchange for higher standards of common knowledge.)

If you are regularly stressed out and scared that you'll be punished by other people if they find out you believe a wrong thing, then either you're overly timid or you're living in an oppressive environment. If fear of punishment or ostracism comes up regularly when you're in the process of forming an opinion, I think that's *too much fear* for critical thinking to work properly at all; and the mantra "I have the right to my opinion" is a good counterweight to that.

Discovery Requires Risking Mistakes

The second sense in which you have a "right to be wrong" is prudential.

You could ensure that you'd *never* be wrong by never venturing an opinion on anything. But going all the way to this extreme is, of course, absurd — you'd never be able to make a decision in your life! The most effective way to accomplish any goal always involves *some* decision-making under uncertainty.

And attempting more difficult goals involves more risk of failure. Scientists make a lot of hypotheses that get falsified; entrepreneurs and engineers try a lot of ideas that don't work; artists make a lot of sketches that wind up in the wastebasket. Comfort with repeated (hopefully low-stakes) failure is *essential* for succeeding at original work.

Even from a purely epistemic perspective, if you want to have the most accurate possible model of some part of the world, the best strategy is going to involve *probabilistically* believing some wrong things; you get information by testing guesses and seeing where you're mistaken. Minimizing error requires finding out where your errors *are*.

Note, though, that from this prudential perspective, it's *not* a good idea to have habits or strategies that systematically bias you towards being wrong. In the "right to your opinion" sense, you have a "right" to epistemic vices, in that nobody should be attacking you for them; but in this goal-oriented sense, they're not going to help you succeed.

Space Mom Accepts All Her Children

The third sense in which you have a "right to be wrong" is a little weirder, so please bear with me.

There's a mental motion you can do, when you're trying to get the right answer or do the right thing, where you're trying very hard to stay on the straight path, and any time you slip off, you violently jerk yourself back on track. You have an *aversion* to wrongness.

I have an intuition that this is…inefficient, or mistaken, somehow.

Instead, there's a mental motion where you have peripheral vision, and you see all the branching paths, and consider where they might go — all of them are possible, all of them are in some cosmic sense "okay" — and you perform some kind of optimization procedure among the paths and then go along the right path smoothly and without any jerks.

Or, consider the space of all mental objects, all possible thoughts or propositions or emotions or phenomena or concepts. Some of these are true statements; some of them are false statements. Most of them are unknown, or not truth-apt in the first place. Now, you don't really want to label the false ones as true — that would just be error. But all of them, true or false or neither or unknown, are *here*, hanging like constellations in this hypothetical heaven. You can look at them, consider them, call some of them pretty. You don't need to have an aversion response to them. They are "valid", as the kids say; even if they don't have the label "true" on them, they're still here in possibility-space and that's "okay".

In a lot of traditions, the physical metaphor for "good" is *high and bright*. Like the sun, or a mountaintop. The Biblical God is described as high and bright, as are the Greek Olympians or the Norse gods; in Indian and Chinese traditions a lot of divine or idealized entities are represented as high and bright; in ordinary English we talk about an idealistic person as "high-minded" and everybody knows that the "light side of the Force" is the side of the good guys.

To me, the "high and bright" ideal feels connected to the pattern of seeking a goal, seeking truth, trying not to err.

But there are also traditions in which "high and bright" needs to be balanced with another principle whose physical metaphor is *dark and vast*. Like the void of space, or the deeps of the sea. Like yin as a complement to yang, or prakrti as a complement to purusa, or emptiness as a complement to form.

The "high and bright" stuff is *value* — knowledge, happiness, righteousness, the things that people seek and benefit from. The "dark and vast" stuff is *possibility*. Room to breathe. Freedom. Potential. Mystery. Space.

You can feel trapped by only seeking value — you can feel like you lack the "space to be wrong". But it's not really that you *want* to be wrong, or that you want the *opposite* of value; what you want is this sense of "enough room to move".

It's something like Keats' "negative capability":

> *…at once it struck me, what quality went to form a Man of Achievement especially in Literature & which Shakespeare possessed so enormously—I mean Negative Capability, that is when man is capable of being in uncertainties, Mysteries, doubts, without any irritable reaching after fact & reason—Coleridge, for instance, would let go by a fine isolated verisimilitude caught from the Penetralium of mystery, from being incapable of remaining content with half knowledge.*

or something like the "mother" Ahab perceives behind God:

> *But thou art but my fiery father; my sweet mother, I know not. Oh, cruel! what hast thou done with her? There lies my puzzle; but thine is greater. Thou knowest not how came ye, hence callest thyself unbegotten; certainly knowest not thy beginning, hence callest thyself unbegun. I know that of me, which thou knowest not of thyself, oh, thou omnipotent. There is some unsuffusing thing beyond thee,*

> *thou clear spirit, to whom all thy eternity is but time, all thy creativeness mechanical. Through thee, thy flaming self, my scorched eyes do dimly see it. Oh, thou foundling fire, thou hermit immemorial, thou too hast thy incommunicable riddle, thy unparticipated grief.*

The womb of nature, the dark vastness of possibility — Space Mom, so to speak — is not the *opposite* of reason and righteousness so much as the *dual* to these things, the space in which they operate. The *opposite* of being right is being wrong, and nobody really wants that per se. The *complement* to being right is something like "letting possibilities arise" or "being curious." Generation, as opposed to selection. Opening up, as opposed to narrowing down.

The third sense in which you have the "right to be wrong" is a lived experience, a way of thinking, something whose slogan would be something like "Possibility Is."

If you have a problem with "gripping too tight" on goals or getting the right answer, if it's starting to get oppressive and rigid, if you can't be creative or even perceive that much of the world around you, you need Space Mom. The impulse to assert "I have the right to disagree even with people who know better than me" seems like it might be a sign that you're suffocating from a lack of Space Mom. You need openness and optionality and the awareness that you *could* do *anything* within your powers, even the imprudent or taboo things. You need to be *free* as well as to be right.

Military AI as a Convergent Goal of Self-Improving AI

This is our accepted chapter in the edited volume Forthcoming as a chapter in Artificial Safety And Security (Roman V. Yampolskiy , ed.), CRC Press.

>Abstract Better instruments to predict the future evolution of artificial intelligence (AI) are needed, as the destiny of our civilization depends on it. One of the ways to such prediction is the analysis of the convergent drives of any future AI, started by Omohundro. We show that one of the convergent drives of AI is a militarization drive, arising from AI's need to wage a war against its potential rivals by either physical or software means, or to increase its bargaining power. This militarization trend increases global catastrophic risk or even existential risk during AI takeoff, which includes the use of nuclear weapons against rival AIs, blackmail by the threat of creating a global catastrophe, and the consequences of a war between two AIs. As a result, even benevolent AI may evolve into potentially dangerous military AI. The type and intensity of militarization drive depend on the relative speed of the AI takeoff and the number of potential rivals. We show that AI militarization drive and evolution of national defense will merge, as a superintelligence created in the defense environment will have quicker takeoff speeds, but a distorted value system. We conclude with peaceful alternatives.

https://www.academia.edu/35130825/Military*AI* as*a* Convergent*Goal* of*Self-Improving* AI

Link without registration and opened for commenting:
https://docs.google.com/document/d/15D71qhhY-ZsAY7syzZsr1lKopTODbdeXVPElaPaIqyA/edit

The Darwin Results

Epistemic Status: True story (numbers are best recollections)

This is post three in the sequence Zbybpu’f Nezl.

Previously (required): The Darwin Game, The Darwin Pregame.

I

It was Friday night and time to play The Darwin Game. Excited players gathered around their computers to view the scoreboard and message board.

In the first round, my score went up slightly, to something like 109 from the starting 100. One other player had a similar score. A large group scored around 98. Others did poorly to varying degrees, with one doing especially poorly. That one played all 3s.

Three, including David, shot up to around 130.

If it isn't obvious what happened, take a minute to think about it before proceeding.

II

The CliqueBots had scores of 98 or so. They quickly figured out what happened.

David lied. He sent the 2-0-2 signal, and cooperated with CliqueBots, but instead of playing all 3s against others, he and two others cooperated with others too.

Whoops.

CliqueBots had been betrayed by MimicBots. The three defectors prospered, and the CliqueBots would lose.

Without those three members, the CliqueBots lacked critical mass. Members would die slowly, then increasingly quickly. If the three defectors had submitted CliqueBots, the CliqueBots would have grown in the first round, reaching critical mass. The rest of us would have been wiped out.

Instead, the three defectors would take a huge early lead, and the remaining members would constitute, as our professor put it, their 'packed lunch.'

The opening consisted of CliqueBots being wiped out, along with G-type weirdos, A-type attackers and D-style cooperators that got zero points from the CliqueBots.

Meanwhile, on the message board, the coalition members were *pissed.*

III

Everyone who survived into the middle game cooperated with everyone else. Victory would come down to efficiency, and size boosted efficiency. Four players soon owned the entire pool: Me and the three defectors.

I thought I had won. The coalition members wasted three turns on 2-0-2. Nothing could make up for that. My self-cooperation was far stronger, and I would outscore

them over the first two rounds when we met due to the 0. It wouldn't go fast, but I would grind them out.

It did not work out that way. David had the least efficient algorithm and finished fourth, but I was slowly dying off as the game ended after round 200. Maybe there was a bug or mistake somewhere. Maybe I was being exploited a tiny bit in the early turns, in ways that seem hard to reconcile with the other program being efficient. I never saw their exact programs, so I'm not sure. I'd taken this risk, being willing to be slightly outscored in early turns to better signal and get cooperation, so that's probably what cost me in the end. Either way, I didn't win The Darwin Game, but did survive long enough to win the Golden Shark. If I hadn't done as well as I did in the opening I might not have, so I was pretty happy.

IV

Many of us went to a class party at the professor's apartment. I was presented with my prize, a wooden block with a stick glued on, at the top of which was a little plastic shark, with plaque on the front saying Golden Shark 2001.

Everyone wanted to talk about was how awful David was and how glad they were I had won while not being him. They loved that my core strategy was so simple and elegant.

I tried gently pointing out David's actions were utterly predictable. I didn't know about the CliqueBot agreement, but I was deeply confused how they didn't see this 'betrayal' coming a mile away. Yes, the fact that they were only one or two CliqueBots short of critical mass had to sting, but was David really going to leave all that value on the table? Even if betraying them hadn't been the plan all along?

They were having none of it. I didn't press. Why spoil the party?

V

Several tipping points could have led to very different outcomes.

If there had been roughly two more loyal CliqueBots, the CliqueBots would have snowballed. Everyone not sending 2-0-2 would have been wiped out in order of how much they gave in to the coalition (which in turn accelerates their victory). Betrayers would have bigger pools, but from there all would cooperate with all and victory would come down to if anyone tweaked their cooperation algorithms to be slightly more efficient. David's betrayal may have cost him the Golden Shark.

If someone had said out loud "I notice that anyone who cares about winning is unlikely to submit the CliqueBot program, but instead will start 2-0-2 and then cooperate with others anyway" perhaps the CliqueBots reconsider.

If enough other players had played more 2s against the CliqueBots, as each of us was individually rewarded for doing, the CliqueBots would have won. If the signal had been 2-5-2 instead of 2-0-2, preventing rivals from scoring points on turn two, that might have been enough.

If I had arrived in the late game with a slightly larger pool, I would have snowballed and won. If another player submits my program, we each end up with half the pool.

Playing more 2s against attackers might have won me the entire game. It also might have handed victory to the CliqueBots.

If I had played a better split of 2s and 3s at the start, the result would have still depended on the exact response of other programs to starting 2s and 3s, but that too might have been enough.

Thus these paths were all possible:

The game ended mostly in MimicBots winning from the momentum they got from the CliqueBots.

It could have ended in an EquityBot (or even a DefenseBot) riding its efficiency edge in the first few turns to victory after the CliqueBots died out. Scenarios with far fewer CliqueBots end this way; without the large initial size boost, those first three signaling turns are a killer handicap.

It could have ended in MimicBots and CliqueBots winning together and dividing the pool. This could happen even if their numbers declined slightly early on, if they survived long enough while creating sufficient growth of FoldBot.

CliqueBots could have died early but sufficiently rewarded FoldBots to create a world where a BullyBot could succeed, and any BullyBot that survived could turn around and win.

It could have had MimicBots and CliqueBots wipe out everyone else, then ended in victory for *very subtle* MimicBots, perhaps that in fact played 3s against outsiders, that exploited the early setup turns to get a tiny edge. Choosing an algorithm that can't be gamed this way would mean choosing a less efficient one.

In various worlds with variously sized initial groups of CliqueBots and associated MimicBots, and various other programs, the correct program to submit might be a CliqueBot, a MimicBot that attacks everyone else but cheats on the coordination algorithm, a MimicBot that also cooperates with others, a BullyBot with various tactics, an EquityBot with various levels of folding, or an FoldBot. There are even scenarios where *all marginal submissions lose,* because the program that would win without you is poisoning the pool for its early advantage, so adding another similar program kills you both.

This is in addition to various tactical settings and methods of coordination that depend on exactly what else is out there.

Everyone's short term interest in points directly conflicts with their long term goal of having a favorable pool. The more you poison the pool, the better you do now, but if bots like you poison the pool too much, you'll all lose.

There is no 'right' answer, and no equilibrium.

What would have happened if the group had played again?

If we consider it only as a game, my guess is that this group would have been unable to trust each other enough to form a coalition, so cooperative bots in the second game would send no signal. Since cooperative bots won the first game, most entries would be cooperative bots. Victory would likely come down to who could get a slight edge during the coordination phase, and players would be tempted to enter true

FoldBots and otherwise work with attackers, since they would expect attackers to die quickly. So there's some chance a well-built BullyBot could survive long enough to win, and I'd have been tempted to try it.

If we include the broader picture, I would expect an attempt to use out-of-game incentives to enforce the rules of a coalition. The rise of a true CliqueBot.

VI

I spent so long on the Darwin Game story and my thinking process about it for several reasons.

One, it's a fun true story.

Two, it's an interesting game for its own sake.

Three, because it's a framework we can extend and work with, that has a lot of nice properties. There's lots to maximize and balance at different levels, no 'right' answer and no equilibrium. It isn't obvious what to reward and what to punish.

Four, it naturally ties your current decisions to your future and past decisions, and to what the world looks like and what situations you will find yourself in.

Five, it was encountered 'in the wild' and doesn't involve superhuman-level predictors. A natural objection to models is 'you engineered that to give the answer you want'. Another is 'let's figure out how to fool the predictor.' Hopefully minimizing such issues will help people take these ideas seriously.

There are many worthwhile paths forward. I have begun work on several. I am curious which ones seem most valuable and interesting, or where people think I will go next, and encourage such discussion and speculation.

Stable agent, subagent-unstable

Crossposted at the intelligent agent forum.

A reflectively consistent agent is one that is ok with creating copies with the same motivation as itself, or continuing to have the same motivation under self-modification.

A reflectively stable agent is one that would only create agents with the same motivation as itself (unless rewarded for doing otherwise), and would actively want to preserve its motivation under self-modification.

Here is a design that is reflectively stable for its own self-modification, but not even reflectively consistent for copies and subagents.

Subtracting the value function

Let h_t be the history of the agent's interaction with the environment up to time t.

Let the agent A have policy π_A. Let $V(R, \pi, h_t)$ be the expected value of (future) reward R under policy π, given history h_t. Let π_R be the policy that maximises this value the most.

Initially, the agent has no reward. Then, at time t, the AI gets reward function R, and gets the penalty:

- $-cV(R, \pi_A, h_t)$.

For $c < 1$, this change of reward is reflexively stable (the agent will choose $\pi_A = \pi_R$ to get maximal expected reward $(1 - c)V(R, \pi_R, h_t)$), for $c = 1$, it is reflexively consistent (any policy π_A gives the same expected reward of 0), and for $c > 1$, it is neither.

But instead, imagine the agent gets the penalty:

- $-cV(R, \pi_R, h_t)$.

In this case, the agent is reflexively stable for all values of c! That's because the penalty term is independent of its policy choice, and it will therefore choose $\pi_A = \pi_R$ after t to maximise the reward it does get from R.

No subagent stability.

Using π_R in the reward may make the agent itself reflexively stable, but if $c > 1$, it will not copy the same goal into its subagent.

This is easy to see, because the agent's expected reward is $(1 - c)V(R, \pi_R, h_t)$, so its best policy is to minimise $V(R, \pi_R, h_t)$. It will do this by acting so that h_t is one the minimises that value.

For instance, it might create subagents whose task is to actively restrain the agent after t. Then it will struggle against them after t (because it now wants to maximise the reward R), but will set them up to be as powerful as possible before that, because before t, it wants its own future struggles to fail.

Clearly, though this is technically reflectively stable, this is not a good sort of stability.

Not 'real' stability

It should be noted that the agent's "stability" is an artefact of the fact that π_R is defined to be "the best policy that the agent can follow to maximise R".

"The agent" is not an ontologically fundamental object, so this stability is only as good as our definition of the agent (just as we could make the whole setup subagent-stable, if only we could define subagents - which we can't really do).

Big Advance in Infinite Ethics

Summary

It is possible that our universe is infinite in both time and space. We might therefore reasonably consider the following question: given some sequences $u = (u_1, u_2, \ldots)$ and $u' = (u'_1, u'_2, \ldots)$ (where each u_t represents the welfare of persons living at time t), how can we tell if u is morally preferable to u'?

It has been demonstrated that there is no "reasonable" ethical algorithm which can compare any two such sequences. Therefore, we want to look for subsets of sequences which can be compared, and (perhaps retro-justified) arguments for why these subsets are the only ones which practically matter.

Adam Jonsson has published a preprint of what seems to me to be the first legitimate such ethical system. He considers the following: suppose at any time t we are choosing between a finite set of options. We have an infinite number of times in which we make a choice (giving us an infinite sequence), but at each time step we have only finitely many choices. (Formally, he considers Markov Decision Processes.) He has shown that an ethical algorithm he calls "limit-discounted utilitarianism" (LDU) can compare any two such sequences, and moreover the outcome of LDU agrees with our ethical intuitions.

This is the first time that (to my knowledge), we have some justification for thinking that a certain algorithm is all we will "practically" need when comparing infinite utility streams.

Limit-discounted Utilitarianism (LDU)

Given $u = (u_1, u_2, \ldots)$ and $u' = (u'_1, u'_2, \ldots)$ it seems reasonable to say $u \geq u'$ if

$$\sum_{t=0}^{\infty} (u_t - u'_t) \geq 0$$

Of course, the problem is that this series may not converge and then it's unclear which sequence is preferable. A classic example is the choice between $(0, 1, 0, 1, \ldots)$ and $(1, 0, 1, 0, \ldots)$. (See the example below.)

LDU handles this by using Abel summation. Here is a rough explanation of how that works.

Intuitively, we might consider adding a discount factor $0 < \delta < 1$ like this:

$$\sum t = 0^{\infty} \delta^{t} (u t - u_{t}')$$

This modified series may converge even though the original one doesn't. Of course, this convergence is at the cost of us caring more about people who are born earlier, which might not endear us to our children.

Therefore, we can take the limit case:

$$\liminf_{\delta \to 1^-} \sum_{t=0}^{\infty} \delta^t (u_t - u_t')$$

This modified summand is what's used for LDU.

LDU has a number of desirable properties, which are summarized on page 7 of this paper by Jonsson and Voorneveld. I won't go into them much here other than to say that LDU generally extends our intuitions about what should happen in the finite case to the infinite one.

Example

Suppose we want to compare $u = (1, 0, 1, 0, \ldots)$ and $u' = (0, 1, 0, 1, \ldots)$. Let's take the standard series:

$$\sum_{i=0}^{\infty} (u_i - u_i') = (1 - 0) + (0 - 1) + (1 - 0) + (0 - 1) + \ldots$$

$$= 1 - 1 + 1 - 1 + \ldots$$

$$= \sum_{i=0}^{\infty} (-1)^i$$

This is Grandi's series, which famously does not converge under the usual definitions of convergence.

LDU though will place in a discount term δ to get:

$$\sum_{i=0}^{\infty} (- 1)^{i} \delta^{i} = \sum_{i=0}^{\infty} (- \delta)^{i}$$

It is clear that this is simply a geometric series, and we can find its value using the standard formula for geometric series:

$$\sum_{i=0}^{\infty} (-\delta)^i = \frac{1}{1+\delta}$$

Taking the limit:

$$\liminf_{\delta \to 1^-} \frac{1}{1+\delta} = 1/2$$

Therefore, the Abel sum of this series is one half, and, since $1/2 > ; 0$, we have determined that $(1, 0, 1, 0, \ldots)$ is better than (morally preferable to) $(0, 1, 0, 1, \ldots)$.

This seems kind of intuitive: as you add more and more terms, the value of the series oscillates between zero and one, so in some sense the limit of the series is one half.

Markov Decision Processes (MDP)

Markov Decision Processes, according to Wikipedia, are:

> At each time step, the process is in some state s, and the decision maker may choose any action a that is available in state s. The process responds at the next time step by randomly moving into a new state s', and giving the decision maker a corresponding reward $R_a(s, s')$.
>
> The probability that the process moves into its new state s' is influenced by the chosen action. Specifically, it is given by the state transition function $P_a(s, s')$.
>
> Thus, the next state s' depends on the current state s and the decision maker's action a.

At each time step the decision-maker chooses between a finite number of options, which causes the universe to (probabilistically) move into one of a finite number of states, giving the decision-maker a (finite) payoff. By repeating this process an infinite number of times, we can construct a sequence $u_1, u_2, \ldots$ where u_t is the payoff at time t.

The set of all sequences generated by a decision-maker who follows a single, time independent, (i.e. stationary) policy is what is considered by Jonsson. Crucially, he shows that **LDU is able to compare any two streams generated by a stationary Markov decision process.** [1]

Why This Matters

My immediate objection upon reading this paper was “of course if you limit us to only finitely many choices then the problem is soluble – the entire problem only occurs because we want to examine infinite things!”

After having thought about it more though, I think this is an important step forward, and MDPs represent an importantly large class of decision processes.

Even though the universe may be infinite in time and space, in any time interval there is plausibly only finitely many states I could be in, e.g. perhaps because there are only finitely many neurons in my brain.

(Someone who knows more about physics than I might be able to comment on a stronger argument: if locality holds, then perhaps it is a law of nature that only finitely many things can affect us within a finite time window?)

Sequences generated by MDPs are therefore plausibly the only set of sequences a decision-maker may need to practically consider.

Outstanding Issues

My biggest outstanding concern with modeling our decisions with an MDP is that the payoffs have to remain constant. It seems likely that, as we learn more, we will discover that certain states are more or less valuable than we had previously thought. E.g. we may learn that insects are more conscious than previously expected, and therefore insect suffering affects our payoffs more highly than we had originally thought. It seems like maybe one could have a “meta-MDP” which somehow models this, but I’m not familiar enough with the area to say for sure.

A more theoretical question is: what sequences can be generated via MDPs? My hope is that one day someone will show LDU (or a similarly intuitive algorithm) can compare any two computable sequences, but I don’t think that this is that proof.

Lastly, we have the standard problems of infinitarian fanaticism and paralysis. E.g. even if our current best model of the universe predicted that MDP was exactly correct, there would still be some positive probability that it was wrong and then our “meta-decision procedure” is unclear.

Conclusion

Overall, I don't think that this completely solves the questions with comparing infinite utility streams, but it's a large step forward. Previous algorithms like the overtaking criterion had fairly "obvious" incomparable streams, with no real justification for why those streams would not be encountered by a decision-maker. LDU is not complete, but we at least have some reason to think that it may be all we "practically" need.

I would like to thank Adam Jonsson for discussing this with me. I have done my best to represent LDU, but any errors in the above are mine. Notably, the justification for why MDP's are all we need to consider is entirely mine, and I'm not sure what Adam thinks about it.

1. This is not explicitly stated in Jonsson's paper, but it follows from the proof of theorem 1. Jonsson confirmed this in email discussions with me.

Against Shooting Yourself in the Foot

Follow-up to: Status Regulation and Anxious Underconfidence

Somehow, someone is going to horribly misuse all the advice that is contained within this book.

Nothing I know how to say will prevent this, and all I can do is advise you not to shoot your own foot off; have some common sense; pay *more* attention to observation than to theory in cases where you're lucky enough to have both and they happen to conflict; put yourself and your skills on trial in every accessible instance where you're likely to get an answer within the next minute or the next week; and update hard on single pieces of evidence if you don't already have twenty others.

I expect this book to be of much more use to the underconfident than the overconfident, and considered cunning plots to route printed copies of this book to only the former class of people. I'm not sure reading this book will *actually* harm the overconfident, since I don't know of a single case where any previously overconfident person was *actually* rescued by modest epistemology and thereafter became a more effective member of society. If anything, it might give them a principled epistemology that actually makes sense by which to judge those contexts in which they are, in fact, unlikely to outperform. Insofar as I have an emotional personality type myself, it's more disposed to iconoclasm than conformity, and inadequacy analysis is what I use to direct that impulse in productive directions.

But for those certain folk who cannot be saved, the terminology in this book will become only their next set of excuses; and this, too, is predictable.

If you were never disposed to conformity in the first place, and you read this anyway… then I won't tell you not to think highly of yourself before you've already accomplished significant things. Advice like that wouldn't have *actually* been of much use to myself at age 15, nor would the universe have been a better place if Eliezer-1995 had made the mistake of listening to it. But you might talk to people who have tried to reform the US medical system from within, and hear what things went wrong and why.[1] You might remember the Free Energy Fallacy, and that it's much easier to save yourself than your country. You might remember that an aspect of society can fall well short of a liquid market price, and still be far above an amateur's reach.

I don't have good, repeatable exercises for training your skill in this field, and that's one reason I worry about the results. But I can tell you this much: *bet on everything.* Bet on everything where you can or will find out the answer. Even if you're only testing yourself against one other person, it's a way of calibrating yourself to avoid both overconfidence and underconfidence, which will serve you in good stead emotionally when you try to do inadequacy reasoning. Or so I hope.

Beyond this, other skills that feed into inadequacy analysis include "see if the explanation feels stretched," "figure out the further consequences," "consider alternative hypotheses for the same observation," "don't hold up a mirror to life and

cut off the parts of life that don't fit," and a general acquaintance with microeconomics and behavioral economics.

The policy of saying only what will do no harm is a policy of total silence for anyone who's even slightly imaginative about foreseeable consequences. I hope this book does more good than harm; that is the most I can hope for it.

For yourself, dear reader, try not to be part of the harm. And if you end up doing something that hurts you: *stop doing it*.

Beyond that, though: if you're trying to do something *unusually well* (a common enough goal for ambitious scientists, entrepreneurs, and effective altruists), then this will often mean that you need to seek out the most neglected problems. You'll have to make use of information that isn't widely known or accepted, and pass into relatively uncharted waters. And modesty is especially detrimental for that kind of work, because it discourages acting on private information, making less-than-certain bets, and breaking new ground. I worry that my arguments in this book could cause an overcorrection; but I have other, competing worries.

The world isn't mysteriously doomed to its current level of inadequacy. Incentive structures have parts, and can be reengineered in some cases, worked around in others.

Similarly, human bias is not inherently mysterious. You can come to understand your own strengths and weaknesses through careful observation, and scholarship, and the generation and testing of many hypotheses. You can avoid overconfidence *and* underconfidence in an even-handed way, and recognize when a system is inadequate at doing *X* for cost *Y* without being exploitable in *X*, or when it is exploitable-to-someone but not exploitable-to-you.

Modesty and immodesty are bad heuristics because even where they're correcting for a real problem, you're liable to overcorrect.

Better, I think, to not worry quite so much about how lowly or impressive you are. Better to meditate on the details of what you can do, what there is to be done, and how one might do it.

This concludes *Inadequate Equilibria*. The full book is now available in electronic and print form through equilibriabook.com.

---

1. As an example, see Zvi Mowshowitz's "The Thing and the Symbolic Representation of The Thing," on MetaMed, a failed medical consulting firm that tried to produce unusually high-quality personalized medical reports. ↩

Lizard Jockeying for Fun and Profit

If this works, it will serve as a kind of introduction to a series - or *sequence*, if you will - on social interaction and strategic incorporation of one's emotions into one's reason. But first let's throw this out there and see what happens, shall we?

Part the zeroth: a caveat.

Let’s face it; with a title like “Lizard Jockeying for Fun and Profit”, this post is promising a wild ride. We’re going to take some appalling metaphoric liberties with modern neuroscience, so bear in mind that we’re dealing in broad, sweeping generalizations throughout most of this paper. Nevertheless, I shall endeavor to back up said sweeping generalizations with something like actual fact.

...Deep breath, now. Here goes!

Part the first: The shape of the monster; or: get to know your lizard before you attempt to mount!

Let's face it: humans love dualities and trinities. For some reason, when we look at ourselves, we really like splitting things into twos and threes whenever we can. Probably because they're very small numbers, and if you're going to be breaking things up to make them easier to understand, you don't want to overwhelm people with too many details.

So, when we get to the ancient Hindu sages, the ancient Greek philosophers, the ancient Austrian psychoanalysts, the ancient American developmental psychologists, or the ancient American neurobiologists, you start seeing tripartate distinctions show up: body, soul, mind; appetite, emotion, reason; id, ego, superego; pre-conventional, conventional, post-convventional; lizard-brain, mammal-brain, ape-brain... it's an easy split.

To tantalize with sparse example: while Lawrence Kohlberg’s stages of moral development are seen as evolutionary rather than competitive, one can draw parallels between the ‘preconventional phase’ (focused on punishment, rewards and immediate personal gain) and the 'id', the ‘conventional phase’ (focused on interpersonal relationships, esteem, and honor) and the 'ego', and the ‘postconventional phase’ (focused on deep principles and rational discourse) and the 'superego'. If we presume that the different portions of our behavior develop from childhood at different rates, and if we accept (as Kohlberg and Piaget themselves did) that the idea of “stages” is itself more of a useful approximation than a hard-and-fast rule, then a picture begins to emerge of Plato's wild, appetitive beast, slowly brought under control by a sense of conviction, and only finally (if we’re lucky) ruled by a fully-developed reason.

All maps are false, but some are true.

Part the second: the shape of the reins; or: checking your equipment before you get on!

Modern evolutionary neurobiology has some insight to shed on where to draw the lines on our map. The human brain, like all biological systems, is a product of millions of years of evolution in mother nature's workshop. And mother nature very, very rarely throws anything away. Systems develop, are used, and then new systems develop on top of them to modify them, those new modified systems are used, and so on.

What this means is that a human being, for all our millions of years of development, still has most of the biological 'hardware' of a lizard. The spinal column, the cerebellum, the pons, the thalamus, the deep amygdala, the medula oblongata - i.e., the deeper parts of our brain - are designed to deal with lizardy things: food and sex and safety and comfort and raw dominance. Not that lizards are the only beings that do this - heck, worms do this - but for the sake of metaphoric imagery and artistic license, let's think of the lizard as the metaphorical "peak" of this kind of brain.

It is here that modern neurobiology can shed some insight on the situation. Many of these organs - the reticular formation, the cerebellum, the pons, and the medula oblongata - have existed at least for the past 500 million years, and developed in early fish. The amygdala developed somewhat later, growing from sensory nerve clusters designed to process smell and pheremone signals. Together, this cluster has first pick of our nerve signals - any information that passes from the body to the rest of the brain must first pass through these centers.

These lizardy bits did a very, very good job at making lizard-like critters the dominant life forms on the planet for millions of years, so there was no real evolutionary pressure for mother nature to ditch them and start over.

So, there we are - the "lizard-brain". You could call it the "id" or the "appetite" or the "pre-conventional morality", but then this paper would have a very silly title for no reason, and we can't have that. Thus, I don the mask of the poet, and "lizard-brain" it is.

Eventually, as our environment became more complex, opportunities to surpass simple lizardy behaviors opened up. And so, the mammalian brain (and other brains to be sure, but mammals do it oh so well) began to develop organs (the temporal and parietal lobes, the insular cortex, and various other systems) and behaviors (herd instincts, empathy, etc.) to reward pack cooperation, to punish defection, to band together against external threats, to bond to individuals and protect them - in short, to love and be loved and to strive for the esteem of the pack.

Many mammals and birds do this exceptionally well - Plato has Socrates use the image of a lion, while a dear friend suggests the noble meerkat - but this is my paper, and I like monkeys. So, in anticipation of future imagery, we will call this section the "monkey-brain". You could call it the "ego" or "honor" or "conventional morality", but that would prevent us from getting to some hilarious mental imagery later. So, poet mask in place, I dub thee "monkey-brain".

The "monkey brain" physically resides within two structures called the the limbic system and the paleocortex. The limbic system contains the hippocampus (which

shares the amygdala with the lizard brain), the limbic lobe, the fornicate gyrus, and the orbitofrontal cortex. These structures, together with the paleocortex, control long-term memory formation, memory-based decision making, and associative learning, and regulate communication between the hindbrain (the “lizard brain”) and the neocortex (the “thinking” brain). They are exceptionally good at associative learning, task learning, and rapid problem-solving, but are still primarily guided by basic drives (pleasure, pain, comfort, social esteem) when deciding what problems to solve.

Now we will descend further into the lurid imagery of metaphor. If I were to cut open your head, we would see a mass of blood and gooey brain tissue - not very useful for evoking the right kind of thought. So instead, let’s imagine the mind from the perspective of our little monkey from the previous paragraph.

Imagine the inside of your head is a giant control room, with a big computer in the back labeled “LIZARD BRAIN - DO NOT OPEN - NO USER SERVICEABLE PARTS INSIDE.” In front of that giant bubbling, liquid-cooled computer is a seat, something like the command chair on a submarine, or Kirk’s chair in the old Star Trek TV show. In front of the seat is a big console full of shiny buttons and levers, and above the console is a screen that shows whatever your eyes are seeing.

In the seat is our monkey.

He has levers in front of him that are labeled “What would your mother think?” and “They’re all going to laugh at you!” and “Chicks dig it!” and “No self-respecting man would do that!”. These are just examples; a less hyper-masculine mind might have a completely different set of labels. But in every single control room, in the very middle of the console, is a big red button labeled “AUTOPILOT - OVERRIDE”.

The monkey’s job is simple. 90% of the time, the Lizard Brain v1.0 computer just steers things around, fulfilling its basic appetitive programing. Sometimes, though, it starts reaching for things that might be bad, or starts running away from things that need to be stood up to. In those moments, the monkey reaches forward and flips a lever, the Lizard Brain thinks (for example) “No self-respecting man would back down from that!”, and the Lizard Brain switches from ‘flight’ mode to ‘fight’ mode.

So you see, the Lizard Brain is doing most of the work; the monkey’s job is just to steer whenever the Lizard Brain 1.0 encounters something beyond its programming. Even when the monkey steers, he’s really just nudging the lizard from one set of lizardy behaviors to another - everything he does relies on the lizard bits working properly in the first place.

Part the third: the lizard-jockey-jockey; or: just when you thought this was easy!

Now, when things get a bit overwhelming, or when the monkey doesn't have the right lever, then the big red "AUTOPILOT - OVERRIDE" button starts flashing, loud warning klaxons go off all over the control room, and the monkey runs around screeching and waving its arms frantically, looking for a lever to pull. Meanwhile, the lizard is roaring and rampaging and making a general mess of things.

There is a modern term for this process, popularized by Daniel Goldman's books on "emotional intelligence" - the amygdala hijack:

"Anatomically the emotional system can act independently of the neocortex. Some emotional reactions and emotional responses can be formed without any conscious, cognitive participation...because the shortcut from thalamus to amygdyla completely bypasses the neocortex."

Now, monkey-brain is pretty clever. Lizard-brain only really thinks about "what": see food, get food, eat food. See girl, get girl, fuck girl. See beer, drink beer, drink more beer, drink more beer, drink even more beer, throw up beer. With monkey-brain in the command seat, "what" can be supplemented with "how" - detailed strategies and plans can be accomplished, and immediate gratification can be delayed for the sake of a better long-term position. But all that assumes that monkey-brain has the right lever to pull, and that lizard-brain is in any mood to pay attention to monkey-brain. Otherwise, we get a rampaging lizard and a screeching, panicked monkey that's going to have to clean up the mess after it's all over.

Of course, even with monkey-brain steering and lizard-brain driving, there are still unexplored vistas of behavior that remain inaccessible to us. Luckily, mother nature is always tinkering, and a few hundred thousand years ago hominids started using all that pack behavior to piggy-back a whole new set of decision-making: the prefrontal cortex. While monkey-brain is good at the clever plans and immediate problem-solving that go into "how", and lizard-brain is exceptionally good at the reward-seeking and pain-avoiding appetites that go into "what", the prefrontal cortex throws a wild new spice into the mix: "why".

For reasons not quite understood, but probably to do with mating behavior, complex courtship displays, and more deception than your average monkey could shake a stick at, at some point our ancestors decided that big brains were sexy. And eventually, we began to use those sexy big brains to create language, and semantics, and narrative structures, and stories, and morality. And then we started using those stories to make something called sense of the world. Like it or not, we're stuck with a sense that "why" matters - and that's why we all have to read Plato, because at some point your ancestors' mating practices led to you being born with a big wrinkled growth sticking out of the front of your brain.

[Ed. note: large chunks of this paper were taken from a writing assignment on Plato. I make no apologies.]

But since we've already established that gooey bits of bloody brain tissue are disgusting, let's paint another metaphor. We go back to the "Mecha-Godzilla Command Room", where our panicked monkey is desperately trying to regain control

over the rampaging lizard brain. If we zoom in further, imagine that we see inside the monkey's head - since this is a metaphorical monkey instead of a real one, we can easily pretend that its head has another tiny control room in it instead of a gooey mass of monkey-brains.

Inside this control room, we imagine, lies a meditating sage. Most of the time, he sits there and watches the monkey do its thing. But sometimes, the monkey lets the lizard do something that the sage doesn't approve of, and the sage must act. He takes the reins of the monkey, and through the monkey, takes the reins of the lizard. Or sometimes, if the lizard doesn't have the right controls, he takes the reins of the monkey and has the monkey build a new lever into the lizard's control room.

All of this, of course, assumes that the monkey lets him. Oftentimes, the monkey likes doing what it wants, and will happily ignore the sage in favor of its own cleverness and self-aggrandizement. When that happens, the sage sighs and watches bemusedly, while the monkey directs the lizard through all sorts of embarrassing monkey behavior - seeking glory and prestige over all the other monkeys with no thought to the greater social fabric.

And suddenly we realize that the monkey has it easy. Every one of those levers that the monkey drives the lizard with got there because someone put it there. Our mom installed the "what would your mother think?" lever before the monkey even learned to pull it. Our dad installed the "no self-respecting man does that!" lever soon after. Our schoolyard friends and foes installed the "they're all going to laugh at you!" lever, whether we want it there or not. By the time we reach adulthood, the monkey's little control panel is just full of levers - assuming he remembers to use them, and assuming they're all wired up right.

The sage, on the other hand, only has one button: "Let's think about this for a minute." The sage has to explain to the monkey what it wants to do, which assumes that the monkey even wants to listen. The sage relies on the monkey's curiosity and cleverness to get the monkey to listen at all, and hopefully, the monkey will agree to install another lever into the lizard.

The neurobiological reasons for this are simple: the neocortex, which is the part of the brain where our reasoning and "higher level" thoughts occur, sits on top of the limbic system. There is no direct connection between it and the "lizard brain"; every communication that the neocortex makes with the rest of our body has to go through that limbic system first, and anything that we sense with our body has to go through the limbic system before the limbic system tells the neocortex about it. Any decision that the neocortex makes has to go through the monkey in order to have an effect on what we actually wind up doing, and any information that the neocortex might want to know about has to be told to it by the limbic system[6].

Part the fourth: free enlightenment or TRIPLE your money back!

So, whenever it seems so difficult to behave like a responsible human being, realize what you're up against. You aren't just a single person with a single set of goals, even though your temporal lobe and your corpus collossum do a bang-up job of making it feel that way. You are a tiny, panicked, stressed-out sage, desperately trying to steer a screeching, poo-flinging, uncooperative little shit of a monkey, in the hopes of using that monkey to steal a giant, terrifying, roaring, rampaging monster of a lizard.

Good luck with that!

Fortunately, you can get better at it. The monkey can install more levers, and can get better at learning which levers are likely to quell which kinds of lizard-rampages. The sage can learn how to talk to the monkey in language the monkey understands. The monkey can learn to listen better to the sage. The lizard can be trained to rampage less often, and to snap out of rampages quickly.

But all this takes work. And the biggest obstacle is the illusion that this work is unnecessary - that in reality, we are already a cohesive whole that operates naturally in an efficient harmony, when nothing could be further from the truth.

Of course, all this can only be described in metaphor, because there is no actual tiny monkey inside your brain, with an actual tiny sage inside the monkey's.

It's only a model. *[Shh!]*

Discussing it without doing it is like having a beautiful map of the world in your house, and never leaving your room. It's like saying, "I don't need to go to Paris, I have this map!"

The map is not the territory. Advice is not experience. And anyone who claims to be able to explain all this in such a way that their explanation is as good as actually working it out for yourself, has no god-damned clue what they are talking about.

XOR Blackmail & Causality

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

[Cross-posted from IAFF.]

I edited my previous post to note that I'm now much less optimistic about the direction I was going in. This post is to further elaborate the issue and my current position.

Counterfactual reasoning is something we don't understand very well, and which has so many free parameters that it seems to explain just about any solution to a decision problem which one might want to get based on intuition. So, it would be nice to eliminate it from our ontology – to reduce the cases in which it truly captures something important to machinery which we understand, and write off the other cases as "counterfactual-of-the-gaps" in need of some other solution than counterfactuals.

My approach to this involved showing that, in many cases, EDT learns to act like CDT because its knowledge of its own typical behavior screens off the action from the correlations which are generally thought to make EDT cooperate in one-shot prisoner's dilemma with similar agents, one-box in Newcomb's problem, and so on. This is essentially a version of the tickle defense. I also pointed out that the same kind of self-knowledge constraint is needed to deal with some counterexamples to CDT; so, CDT can't be justified as a way of dealing with cases of failure of self-knowledge in general. Instead, CDT seems to improve the situation in some cases of self-knowledge failure, while EDT does better in other such cases.

This suggests a view in which the self-knowledge constraint is a rationality constraint, so the tickle defense is thought of as being true for rational agents, and CDT=EDT under these conditions of rationality. I suggested that problems for which this was not true had to somehow violate the ability of the agent to perform experiments in the world; IE, the decision problem would have to be set up in such a way as to prevent the agent from decorrelating its actions from things in the environment which are not causally downstream of its actions. This seems in some sense unfair, as the environment is preventing the agent from correctly learning the causal relationships through experimentation. I called this condition the law of logical causality when it first occurred to me, and mixed-strategy implementability in the setup where I proved conditions for CDT=EDT.

In XOR Blackmail with a perfect predictor, however, mixed-strategy implementability is violated in a way which does not intuitively seem unfair. As a result, knowledge of what sort of thing you do in XOR blackmail is not sufficient to decorrelate your actions from things which you have no control over. Constraining to the epsilon-exploration case, so that conditional probabilities are well-defined, it seems like what happens is that the epsilon-exploration bit correlates the action you take with the disaster (thanks to the XOR which determines if the letter is sent). On the other hand, it seems as if CDT should be able to get the right answer.

However, I'm unable to come up with a causal Bayes net which seems to faithfully represent the problem, so that I can properly compare how CDT and EDT reason about it in the same representation. It seems like the letter has to be both a parent and a child of the action. I thought I could represent things properly by having a copy of the

action node, representing the simulation of the agent which the predictor uses to predict; but, I don't see how to represent the perfect correlation between the copy and the real action without effectively severing the other parents of the real action.

Anyone have ideas about how to represent XOR Blackmail in a causal network?

Edit:

Here we go. I was confused by the fact that CDT can't reason as if its action makes the letter not get sent. The following causal graph works well enough:

Variables:

- ***A***: the action. True if money is sent to the blackmailer.
- ***A´***: a copy of the action, representing the abstract mathematical fact of what the agent does if it sees the letter.
- ***L***: Whether the letter is sent or not.
- ***D***: The rare disaster.
- ***U***: The utility.

Causal connections:

- ***A***: Has ***A´*** and ***L*** as parents, with the following function: If the letter is sent, copy ***A´***. Otherwise, false.
- ***A´***: No parents.
- ***L***: ***A´*** and ***D*** as parents, with the XOR function determining ***L***.
- ***D***: No parents.
- ***U***: ***A*** and ***D*** as parents, with the utility function as stated in the original XOR post.

Assume epsilon-exploration to ensure that the conditional probabilities are well-defined. Even if EDT knows its own policy, it sees itself as having control over the disaster. CDT, on the other hand, sees no such connection, so it refuses to send the money.

Remove Intercom?

I posted a question there (about where to find the RSS feeds) and nobody answered. It looks like nobody else has posted there either. So apparently this widget is no longer used, and questions should be asked in some other way?

The Darwin Pregame

Epistemic Status: True story

This is intended as post two of the sequence Zbybpu'f Nezl.

Previously (required): The Darwin Game

I

This is my reconstruction of my thoughts at the time.

The Darwin Game requires surviving the early, middle and late games.

In the opening, you need to maximize scoring against whatever randomness people submit. Survival probably isn't enough. The more copies of yourself you bring to the middle game, the more you face yourself, which snowballs. Get as many points as you can.

In the middle game, you face whatever succeeded in the opening. Strategies that survived the opening in bad shape can make a comeback here, if they are better against this new pool. What strategies do well *against you* matters.

In the end game, you'll need to beat the successful middle game strategies, all of which have substantial percentages of the pool. Eventually you'll be heads up against one opponent. Not letting opponents outscore you in a pairing becomes vital.

How would the game play out? What types of strategies would thrive?

I divided the types as follows:

There were attackers, who would attempt to get the opponent to accept a 3/2 or 4/1 split. They might or might not give up on that if you refused, and presumably most would use a signal to self-cooperate, but not all. One person did submit "return 3."

Then there were cooperators, who attempt to split the pot evenly. I assumed that meant alternating 3/2 splits. This then divided into those who would fold if attacked, allowing you to score above 2.5 per turn, those that would let themselves be outscored but would make sure you scored less than 2.5 per turn, and those that would not allow themselves to be outscored. The last group might or might not forgive an early attempt to attack them.

There would also be bad programs. People do dumb things. Someone might play all 2s, or pick numbers fully at random, or who knows what else.

As a list (attackers from here on means both AttackBot and BullyBot):

AttackBot. Attackers who don't give up.

BullyBot. Attackers who give up.

CarefulBot. Cooperators who harshly punish attackers.

DefenseBot. Cooperators who don't let you outscore them but don't otherwise punish.

EquityBot. Cooperators who let you outscore them, but make sure you don’t benefit.

FoldBot. Cooperators who accept full unfavorable 3/2 splits.

GoofBot. Weird stuff.

My prior was we’d see all seven, with most looking to cooperate.

Was attacking a good strategy?

Attacking only works against FoldBots. When attacking fails, even DefenseBots might take a while to re-establish cooperation. CarefulBots could wipe you out. It was also impossible to know how long to keep attacking before concluding opponents weren’t going to fold.

With a pool of bots chosen by humans, attacking strategies (AttackBot or BullyBot) likely would fail hard in the opening.

The endgame was a different story. All GoofBots would be dead. Unless FoldBots fold too quickly to a BullyBot, in a given round they strictly outscore CarefulBots, DefenseBots and EquityBots. Each round, provided they exist, FoldBots would become a bigger portion of the cooperative pool. If you were an AttackBot or BullyBot, and survived long enough, you would kill off the CarefulBots, then the DefenseBots and finally the EquityBots as the FoldBots out-competed them, leaving a world of AttackBots, BullyBots and FoldBots. If all but one attacker was gone, the last attacker to survive would win if it cooperated efficiently against itself, since it would score above average each round. In theory a steady state could exist with multiple attackers keeping each other in check, but that isn’t stable since advantages in size snowball.

CarefulBots are strictly worse than DefenseBots, so those were out. GoofBots are terrible.

This meant there were five choices:

I could submit an AttackBot that cooperates with itself, and hope to survive into the endgame. I quickly dismissed this as unlikely to work.

I could submit a BullyBot that cooperates with itself, attacks but accepts an even split against stubborn opponents. But this rewards stubborn opponents while wiping out non-stubborn opponents in the mid-game, which means your endgame trump card stops working. I dismissed this as well.

DefenseBots don’t lose heads-up by non-tiny amounts, and punish anyone who tries to outscore them, wiping them out in the mid-game. But you score nothing against AttackBots in the opening, before you can shape the pool much. At best you take a smaller pool into the mid-game, where efficient cooperation with your own copies starts to snowball.

I saw the *emotional* appeal of DefenseBots, but using one didn’t made sense. Its defenses were too robust and expensive, and you still lose to a smart AttackBot heads-up if you’re outnumbered. I’d need to take more risk.

That was the problem with being a FoldBot. FoldBots feed attackers. You are free riding on the rest of the cooperative pool. You hope they kill attackers despite that. The problem is that if even one copy of an attacker survives, as you and other

FoldBots grow strong, attacking becomes a better and better strategy. I decided this wasn't worth that risk.

I would submit an EquityBot. I wouldn't protect against them *outscoring me.* I would protect against them *outscoring what cooperation would have gotten them.* If at any point they wanted to split the remaining pie, I would accept. Even if they refused, I'd give them *some* points on a 3/2 split, so long as they were punished for it, and I wasn't growing their portion of the pool.

This raised the threshold percentage of the pool I needed to win heads-up against an attacker, but with a size disadvantage I'd lose no matter what, and I'd still win if I had a sufficiently large size edge, which was more likely if I did better early on.

Too much folding and you strengthen someone who beats you. Too little and you fall behind letting others snowball.

I decided to alternate 3/2 even if my opponent was going 3/3. This said both 'I'm not going to give up' and 'you are welcome to cooperate at any time,' and still punished the opponent reasonably hard. After long enough I even risked throwing in a few more 2s.

I considered sending a signal to recognize myself, but realized there was no point. Better to start coordinating right away. I'd randomize my first turn to 2 or 3, and once my opponent didn't match me I would alternate. I figured opponents would start 2 more often than 3, so I decided to do a 50/50 split to take advantage of that, coordinating faster and with a slight edge, at the expense of doing slightly worse against myself, but this was probably just a mistake and I should have done an uneven split (but not quite the fully maximizing-for-self-play ratio). However, in an endgame against a similar program, you can definitely get an edge by being slightly more willing to play 3s early than your opponent.

Opponents that wanted to cooperate would have a very easy time recognizing my offer and cooperating. That left special case logic.

If my opponent was alternating on the same schedule as me (somehow we started 2/3, but then we'd 2/2 then 3/3 then 2/2), then I'd play 2 twice in a row to break that up. Ideally, if the opponent was offering a different cycle that was fair, I'd match that (so if they went 1/4/1/4, I'd submit 4 next time, and if they did 1 I'd start alternating), but I didn't expect such cases so I didn't make that logic robust, as the professor had already thrown out part of a previous submission for being too complex, and I wanted to preserve the more important parts.

If my opponent was playing all 2s even after I started alternating, I put in logic to play all 3s. If they played even one 3, I'd back down permanently. I also put in logic against a few other bizarre simple bots (like all 1s, all 4s, seems to be completely random, etc) but didn't worry about it too much since they'd be wiped out very quickly and complexity is bad.

If my opponent was playing all 3s without a starting signal, and kept it up long enough, that meant he'd defect against himself, which meant he couldn't win an endgame, and also meant that he was highly unlikely to ever give up, so I'd eventually fold. If they were going to lose in the long run, better to get what I could. Letting them survive longer would only help me.

II

David took a different approach.

David knew about the class mailing list.

David assembled a large group. They agreed to submit 2-0-2 as their first three moves. If both sides sent the signal, they'd cooperate using a reasonable randomization system. If they didn't get the signal back, they'd play all 3s. They'd be pure CliqueBots, cooperating with each other and defecting against everyone else. With a large enough group, they'd wipe out the other players and share the victory. David would win The Golden Shark and his guaranteed A+.

I would find out about the coalition after round one.

III

We were all set for game night. We had each chosen the logical output of our decision functions. The professor set up a website where we could see the game played out in real time over the course of several hours (due to a combination of that's more fun and the game was slow to run), with a discussion board for him to offer observations and us to comment.

Next time I'll reveal what happened on game night. Predictions are encouraged.

USA v Progressive 1979 excerpt

In 1979, an interesting judgment was made regarding the publication of an alleged nuclear infohazard. Here is an excerpt from that preliminary injunction ruling, which was authored by Robert W. Warren, then a Wisconsin Eastern District judge.

"The Secretary of State states that publication will increase thermonuclear proliferation and that this would "irreparably impair the national security of the United States." The Secretary of Defense says that dissemination of the Morland paper will mean a substantial increase in the risk of thermonuclear proliferation and lead to use or threats that would "adversely affect the national security of the United States."

Howard Morland asserts that "if the information in my article were not in the public domain, it should be put there . . . so that ordinary citizens may have informed opinions about nuclear weapons."

Erwin Knoll, the editor of The Progressive, states he is "totally convinced that publication of the article will be of substantial benefit to the United States because it will demonstrate that this country's security does not lie in an oppressive and ineffective system of secrecy and classification but in open, honest, and informed public debate about issues which the people must decide."

The Court is faced with the difficult task of weighing and resolving these divergent views.

A mistake in ruling against The Progressive will seriously infringe cherished First Amendment rights. If a preliminary injunction is issued, it will constitute the first instance of prior restraint against a publication in this fashion in the history of this country, to this Court's knowledge. Such notoriety is not to be sought. It will curtail defendants' First Amendment rights in a drastic and substantial fashion. It will infringe upon our right to know and to be informed as well.

A mistake in ruling against the United States could pave the way for thermonuclear annihilation for us all. In that event, our right to life is extinguished and the right to publish becomes moot.

In the Near case, the Supreme Court recognized that publication of troop movements in time of war would threaten national security and could therefore be restrained. Times have changed significantly since 1931 when Near was decided. Now war by foot soldiers has been replaced in large part by war by machines and bombs. No longer need there be any advance warning or any preparation time before a nuclear war could be commenced.

In light of these factors, this Court concludes that publication of the technical information on the hydrogen bomb contained in the article is analogous to publication of troop movements or locations in time of war and falls within the extremely narrow exception to the rule against prior restraint.

Because of this "disparity of risk," because the government has met its heavy burden of showing justification for the imposition of a prior restraint on publication of the objected-to technical portions of the Morland article, and because the Court is

unconvinced that suppression of the objected-to technical portions of the Morland article would in any plausible fashion impede the defendants in their laudable crusade to stimulate public knowledge of nuclear armament and bring about enlightened debate on national policy questions, the Court finds that the objected-to portions of the article fall within the narrow area recognized by the Court in Near v. Minnesota in which a prior restraint on publication is appropriate.

The government has met its burden under section 2274 of The Atomic Energy Act. In the Court's opinion, it has also met the test enunciated by two Justices in the New York Times case, namely grave, direct, immediate and irreparable harm to the United States.

The Court has just determined that if necessary it will at this time assume the awesome responsibility of issuing a preliminary injunction against The Progressive's use of the Morland article in its current form."

The case never got to the Supreme Court because the relevant technical details were, in the intervening time, published by others.

More in:

- The full USA vs Progressive preliminary injunction ruling
- USA vs Progressive on Wikipedia

Creating Welfare Biology: A Research Proposal

[This idea came out of an ACE research workshop. I would like to thank Zach Groff and Mark Budolfson for brainstorming this idea with me, as well as ACE for offering me the opportunity to think of it.]

[Crossposted with no modifications from my blog.]

Many people in the wild-animal suffering space think it would be a good idea to make a discipline of "welfare biology"-- that is, the scientific study of wild animal welfare, the way that animal welfare studies scientifically studies domestic animal welfare. From my perspective, there are two big benefits to creating welfare biology. First, it would probably increase the number of research dollars that go to wild-animal welfare research, while reducing the opportunity cost: welfare biology would be competing with other fields of biology for funding, not with starving Africans and tortured pigs. Second, it would give us academic credibility. In most of the developed world, terrestrial wildlife often live on government land (for example, much of the United States's wildlife lives on the quarter of US land owned by the government), which means changing government policies towards wildlife is a promising method of improving their welfare. Even in human-inhabited areas, changing government policies may be an effective way of improving wild-animal welfare. Governments are generally more likely to listen to tenured academics than they are to bloggers.

However, it is unclear to me how one creates an academic field. It is possible that people already know how academic fields form; I have not studied the subject in depth, and would welcome links from commenters. But if there is not already an academic body of work on the subject then it seems useful to do a small research project to explore how academic fields form. I think the best method is a series of qualitative case studies exploring how various relevant scientific fields formed.

I'm aware of two similar research projects in the effective altruist community. Luke Muehlhauser has written a report on early field growth. However, his report concentrates on the role of philanthropists and only touches on what non-philanthropists can do. It also mostly examines fields relevant to artificial intelligence risk research, which has some overlap with fields relevant to welfare biology but not entirely. Animal Charity Evaluators has done several case studies of other social movements; however, its case studies focus on social movements more broadly rather than academic fields specifically.

Histories of academic fields don't usually have the information I'd want from them. For example, this paper-- a fairly typical history of conservation biology-- highlights several important milestones, but doesn't talk about the details. There's a lot of emphasis on particular research projects and controversies, such as whether large reserves are better than small reserves, but not a lot of nitty-gritty detail about how so many ecologists became interested in conservation and how they created common knowledge that they were all interested in them. Nevertheless, the histories do identify key events and key historical figures.

Many academic fields have formed relatively recently; it makes sense to study recently-formed fields, because academia changes over time. So most of the key

historical figures involved in forming a particular academic field may still be alive. The researcher can contact these figures and set up qualitative interviews, focusing on the information that gets left out of histories. How did people get interested in the topic? How did they meet other people who were interested? What steps (journals, book publications, conferences, something else I haven't thought of) were particularly important in getting the field to be self-sustaining? The researcher should also ask for recommendations of sources written at the time and thus supplement their interviews with archival research (to help compensate for the interviewees' poor or self-serving memories).

Once several case studies have been conducted, the researcher can look for common themes. For example, perhaps popular attention or an activist movement galvanized academics into forming the field. Perhaps founding a journal gave people a place to submit papers that otherwise would have languished, unsuited for any currently existing journal. Perhaps an exciting book publication got everyone talking about the subject. This can inform wild animal advocates' strategies in forming welfare biology.

No field is going to be exactly analogous to welfare biology; no one has ever attempted to scientifically study how humans can improve the well-being of wild animals before (although the study of animal emotions explores wild animals' well-being more generally). However, there are several characteristics that might make a field more analogous. Welfare biology is value-laden: that is, instead of just collecting facts about the world, welfare biology is intended to *change* the world. People who think wild-animal suffering is just peachy are unlikely to be interested in welfare biology, just as people who don't care about environmental preservation probably don't care about conservation biology. The researcher might want to emphasize interdisciplinary fields, particularly fields that grew out of biology or that focus on animals. The researcher will probably want to avoid purely social-science or humanities fields, which may not be generalizable to natural science.

Fields it may be interesting to study include:

- Conservation biology (possibly the field most analogous to welfare biology, as a value-laden science that focuses on wild animals and plants)
- Animal welfare science (grew out of attempts to maximize productivity, thus may not be applicable)
- Environmental engineering
- Artificial intelligence risk (interesting because it is a field in the process of forming; advocates for future people, who like animals cannot self-advocate)
- Environmental ethics
- Environmental economics
- Developmental economics (focuses on people in the developing world, who have a limited ability to advocate for themselves in the developed world)
- Positive psychology

It is important to do case studies of failed attempts to start a field, as well as successful ones. Some of the apparent common ground between successful attempts might actually be common ground between all attempts, whether successful or unsuccessful. However, it is much more difficult to find a list of failed attempts than it is to find a list of successful attempts. I suggest that the researcher should ask their early interviewees for ideas, since the interviewees might have personally witnessed unsuccessful attempts to start an academic field. Luke Muehlhauser's report on early field growth briefly discusses cryonics and molecular nanotechnology, which may be interesting fields to review.

I will not personally be able to perform this research project, but I think it’s an interesting and important project for someone to take up, and I’d be happy to consult with any effective altruists who want to do it. Ideally, the researcher would have expertise in qualitative research, particularly interviewing. The final report could create knowledge useful not only for wild-animal advocates but for any effective altruists who want to create an academic field.

Our values are underdefined, changeable, and manipulable

Crossposted at Intelligent Agent Forum.

When asked whether "communist" journalists could report freely from the USA, only 36% of 1950 Americans agreed. A follow up question about Amerian journalists reporting freely from the USSR got 66% agreement. When the order of the questions was reversed, 90% were in favour of American journalists - and an astounding 73% in favour of the communist ones.

There are many examples of survey responses depending on question order, or subtle issues of phrasing.

So there are people whose answers depended on question order. What then are the "true" values of these individuals?

Underdetermined values

I think the best way of characterising their values is to call them "underdetermined". There were/are presumably some people for which universal freedom of the press or strict national security were firm and established values. But for most, there were presumably some soft versions of freedom of the press and nationalism, and the first question triggered one narrative more strongly than the other. What then, are their "real" values? That's the wrong question - akin to asking if Argentina really won the 1986 world cup.

Politicians can change the opinions of a large sector of the voting public with a single pronouncement - were the people's real opinions the ones before, or the ones after? Again, this seems to be the wrong question. But don't people fret about this inconsistency? I'd wager that they aren't really aware of this, because people are the most changeable on issues they've given the least thought to.

And rationalists and EAs are not immune to this - we presumably don't shift much on what we identify as our core values, but on less important values, we're probably as changeable as anyone. But such contingent values can become very strong if attacked, thus becoming a core part of our identity - even if it's very plausible we could have held the opposite position in a world slightly different.

Frameworks and moral updating

People often rely on a small number of moral frameworks and principles to guide them. When a new moral issue arises, we generally try and fit it into a moral framework - and when there are multiple ones that could fit, we can go in multiple directions, driven by mood, bias, tribalism, and many other contingent factors.

The moral frameworks themselves can and do shift, due to issues like tribalism, cognitive dissonance, life experience, and our own self-analysis. Or the frameworks can accumulate so many exceptions or refinements, that they transform in practice if

not in name - it's very interesting that my leftist opinions agree with Anders Sandberg's libertarian opinions on most important issues. We seem to have changed positions without changing labels.

Metaethics

In a sense, you could see all of metaethics as the refinement and analysis of these frameworks. There are urges towards simplicity, to get a more stable and elegant system, and towards complexity, to capture the full spectrum of human values. Much of philosophical disagreement can be seen as "Given A, proposition B (generally acceptable conclusion) implies C (controversial position I endorse)", to which the response is "C is wrong, thus A (or B) is wrong as stated and needs to be refined or denied" - the logic is generally accepted, but which position is kept varies.

Since ethical disagreements are rarely resolved, it's likely that the positions of professional philosophers, though more consistent, are also often driven by contingent and random factors. The process is not completely random - ethical ideas that are the least coherent, like the moral foundation of purity, tend to get discarded - but is certainly contingent. As before, I argue you should focus on the procedure P by which philosophers update their opinions, rather than the (hypothetical) R to which P may be supposed to converge to.

Most people, however, will not have consistent meta-ethics, as they haven't considered these questions. So their meta-opinions there will be even more subject to random influences that their base-level opinions.

Future preferences

There is an urgent question dividing the future world: should local FLOOBS be allowed to restrict use of BLARGS, or instead ORFOILS should pressure COLATS to agree to FLAPPLE the SNARFS.

Ok, we don't currently know what future political issues will be, but it's clear there will be new issues (how do we know this? Because nobody cares today whether Richard Lionheart and Phillip August of France lacked in their feudal duties to each other, nor did the people of that period worry much about medical tort reform). And people will take positions on them, and they will be incorporated into moral frameworks, causing those frameworks to change, and eventually philosophers may incorporate enough change into new metaethical frameworks.

I think it's fair to say that our current positions on these future issues are even more under-determined than most of our values.

Contingent means manipulable

If our future values are determined by contingent facts, then a sufficiently powerful and intelligent agent can manipulate our values, by manipulating those facts. However, without some sort of learning-processes-with-contingent-facts, our values are underdetermined, and hence an agent that wanted to maximise human values/reward wouldn't know what to do.

It was this realisation, that the agent could manipulate the values it was supposed to maximise, that caused me to look at ways of avoiding this.

Choices need to be made

We want a safe way to resolve the under-determination in human values, a task that gets more and more difficult as we move away from the usual world of today and into the hypothetical world that a superpowered AI could build.

But, precisely because of the under-determination, there are doing to be multiple ways of resolving this safely. Which means that choices will need to be made as to how to do so. The process of making human values fully rigorous, is not value-free.

(A minor example, that illustrated for me a tiny part of the challenge: does the way we behave when we're drunk reveal our true values? And the answer: do you want it to? If there is a divergence in drunk and sober values, then accommodating drunk values is a decision - one that will likely be made sober.)

The Journal of High Standards

Our current academical journals are horrible. On the one hand charge a lot of money for access to their papers. On the other hand the fail to uphold the research standards they endorse.

The New England Journal of Medicine for example has endorsed the CONSORT guidelines on best practice in trial reporting. When Ben Goldacre sent him a letter about how out of 23 papers Ben Goldacre analysed only 3 were completely in accord with the CONSORT guidelines, they weren't even willing to print the letter.

The CONSORT guidelines don't require that researchers do fancy statistics instead of using p-values but just require basics like a paper explicitly mentioning when it reports outcomes that weren't preregistered. If the journal would care about high scientific standards it wouldn't be a big deal to uphold the CONSORT guidelines. Unfortunately, the other top medical journals aren't much better.

Why is it hard to establish a new journal that has higher standards? As Yudkowsky writes in Moloch's toolbox that academics have to publish in journals of high prestige and journals happens to be in high prestige when high quality papers are submitted to them. It's a chicken-and-egg problem.

How could we start a new journal that would get scientists willing to submit papers to the journal? If the journal would pay every author of a paper $50,000 for publishing the paper scientists would have an incentive to publish in the journal.

Even if the money alone isn't enough to warrant the scientist to publish in a no-name journal, the journal would soon stop being a no-name journal because scientists would expect that their colleges want to publish in the journal to get the money. That expectation makes the journal more prestigious. The expectations that other people expect the journal to get more prestigious in-turn will increase it's prestige.

Funding this Journal of High Standards wouldn't be a cheap project, but the money that payed to the scientists who publish the papers isn't going to waste. It's like giving a grant to the scientist expect that when you give them the grant you usually ask them what they want to do with the grant money and in this system they can decide for themselves what they are going to do with the money.

It would be a bit like XPrizes but instead of giving the scientists specific goals, it's "Do something worthwhile and then write a high quality paper about it".

Anthropic reasoning isn't magic

The user Optimization Process presented a very interesting collection of five anthropic situations, leading to seemingly contradictory conclusions where we can't conclude anything about anything.

It's an old post (which I just discovered), but it's worth addressing, because it's wrong - but very convincingly wrong (it had me fooled for a bit), and clearing up that error should make the situation a lot more understandable. And you don't need to talk about SSA, SIA, or other advanced anthropic issues.

The first example is arguably legit; it's true that:

> And so, a universe created by some kind of deity, and tuned for life, is indistinguishable from a universe that happens to have the right parameters by coincidence. "We exist" can't be evidence for our living in one or the other, because that fact doesn't correlate with design-or-lack-of-design

But what's really making the argument work is the claim that:

> unless you think that, from *inside a single universe*, you can derive sensible priors for the frequency with which *all* universes, both designed and undesigned, can support life?

But the main argument fails at the very next example, where we can start assigning reasonable priors. It compares worlds where the cold war was incredibly dangerous, with worlds where it was relatively safer. Call these "dangerous", and "safe". The main outcome is "survival", ie human survival. The characters are currently talking about surviving the cold war - designate this by "talking". Then one of the character says:

> **Avery**: Not so! Anthropic principle, remember? If the world had ended, we wouldn't be standing here to talk about it.

This encodes the true statement that P(survival | talking) is approximately 1, as are P(survival | talking, safe) and P(survival | talking, dangerous). In these conditional probabilities, the fact that they are talking has screened off any effect of the cold war on survival.

But Bayes law still applies, and

- P(dangerous | survival) = P(dangerous) * (P(survival | dangerous)/P(survival)).

Since P(survival | dangerous) < P(survival) (by definition, dangerous cold wars are those where the chance of surviving are lower than usual), we get that

- P(dangerous | survival) < P(dangerous).

Thus the fact of our survival has indeed caused us to believe that the cold war was safer than initially thought (there are more subtle arguments, involving near-misses, which might cause us to think the cold war was more dangerous than we thought, but those don't detract from the result above).

The subsequent examples can be addressed in the same way.

Even the first example follows the same pattern - we might not have sensible priors to start with, but if we did, the update process proceeds as above. But beware - "a deity constructed the universe specifically for humans" is strongly updated towards, but that is only one part of the more general hypothesis "a diety construced the universe specifically for some thinking entities", which has a much weaker update.

What is anthropic reasoning for, then?

Given the above, what is anthropic reasoning for? Well, there are sublte issues with SIA, SSA, and the like. But even without that, we can still use anthropic reasoning about the habitiability of our planet, or about the intelligence of creatures realted to us (that's incidentally why the intelligence of dolphins and octopuses tells us a lot more about the evolution of intelligence, as they're not already "priced in" by anthropic reasoning).

Basically, us existing does make some features of the universe more likely and some less likely, and taking that into account is perfectly legitimate.

Confessions of a Slacker

Crossposting the entire thing from Putanumonit in honor of the Slack sequence.

Go read Slack, it's short and important.

If you're not a slacker, that post might change your life and explain why you feel like you have no control over your own life despite doing well on *almost* all counts. If you are a slacker, like me, this post gives our philosophy a name and provides a definition: *slack is the absence of binding constraints on behavior*.

Zvi's post is abstract on purpose. I'll continue his mission by getting more specific and, of course, by putting a num on it. To the latter purpose, I'll modify the definition of slack to make it quantifiable:

Slack is the *distance from* binding constraints on your behavior.

Keep your Distance

Slack is a function of many resources. Running out of any single vital resource is enough to constrain your behavior: make you do something you didn't want to, or prevent you from doing something you want. Freedom requires having spare time, spare money, spare energy, spare weirdness points, available friends etc. The "slack as distance" formula looks a little something like this:

$$Scarcity = \frac{1}{Slack} = \sqrt{\left(\frac{1}{spare\ time}\right)^2 + \left(\frac{1}{spare\ money}\right)^2 + \left(\frac{1}{spare\ energy}\right)^2 + \cdots}$$

Slack disappears when the spare capacity of any single resource goes to zero, regardless of how much of everything else you have. Maintaining slack requires balancing all the important resources, making sure to shore up the scarcest resources first.

My grandma just paid to replace a pipe on her floor that was flooding the entire apartment building. The other 20 tenants were supposed to participate in the cost, but due to diffusion of responsibility and greed, they decided collectively to weasel out of contributing. A lawyer suggested that my grandma should go to court but she refused, for reasons of slack. The question isn't whether the time in court will be worth the money *gained*, but whether the *lack* of this particular sum of money will *force* my grandma into something as undesirable as spending weeks litigating against her neighbors. It almost certainly wouldn't.

At her age, my grandma's scarcest resource is stress-free time. She's not trading it for money.

It's remarkable how many people fuck up by doing the opposite: concentrating on the resources that are easiest for them to obtain, and neglecting their most pressing needs.

One half of Americans have less than $400 to spare, including millions of middle-class people who make and spend tens of thousands each year. The author of this shocking article is an educated professional and a family man. He has accumulated many achievements, but he forgot to save any cash to pay the water bill. On the other side, I have friends from business school with six figures to spare in their bank account and not a single hour to relax.

Whatever resource is scarcest for people is probably the one they aren't good at dealing with, and for this reason, thinking about it is aversive. It's easier just to ignore it. But an ignored constraint doesn't go away, it still binds you.

Currency Exchange

When no single resource is very scarce, you can keep it that way by figuring out the exchange rates between resources and making trade-offs based on those. Trading-off "life currencies" is a subject I discussed in detail before, but slack-based thinking offers a good way to calculate the correct exchange rates. For example: how much is an hour of free time worth to you in dollars?

The slack-based exchange rate is [spare money] / [spare time], when the definition of "spare" is derived from the definition of slack. **Spare X** = how much X you **can lose before being forced** into undesirable behavior.

Example: you make $60,000 a year, put $15,000 in retirement, save $5,000 in cash and spend the remaining $40,000. Of those, $30,000 are for necessities and $10,000 are for things you can live without like fancy clothes and expensive restaurants. This means that you can cover your necessities for $45,000 and since you make $60,000, that means you have $15,000 a year in spare cash. That's how much money you can give up before being forced to change your lifestyle significantly (e.g. move to a cheaper apartment) or jeopardizing your retirement.

Do the same math for time spent: let's say you spend 10 hours a week on activities other than those you *have* to do (work, sleep) or those you really *want* to do (ping pong). These 10 hours a week (or ~500 a year) aren't the entirety of your free time, they're the hours you can afford to lose without having to sacrifice important activities.

In our example, $15,000 spare money each year and 500 spare hours imply an exchange rate of $30/hour. This is a good baseline to consider trade-offs against.If you can pay a maid service $75 to save you three hours of house cleaning ($25/hour), you should take the opportunity because you're converting money to time at a good exchange rate.

Notice that once you've made a trade-off, the exchange rate shifts. If the cleaner comes once a month it saves you 36 hours each year and costs you $900. You now have $14,100 to spare and 536 hours, so the new implied exchange rate is 14100/536 = $26.3/hour. Money became scarcer relative to free time, and you'll be less inclined to keep trading it away.

One danger that lurks when calculating the trade-offs is forgetting about the important resources that are hard to measure. A while after my friend hired a housekeeper, his girlfriend remarked that if they had done this earlier she probably would have had a lot more sex with him. What's the resource that was binding the girlfriend before the housekeeper showed up? I don't think it's spare time or even energy. If you spend your last hour and ounce of energy dusting shelves instead of making love, slack isn't the problem in your relationship.

So what is it? As usual, we shall find the answer in the ancient teaching of the Hebrew sages.

Abraham and Three Angels, Fiasella Domenico

The only time in the Old Testament when God tells a bald-faced lie is when informing Abraham (age 99) and Sarah (88) of their upcoming pregnancy. When Sarah hears the news she laughs incredulously, wondering how can she have a son when *"... my husband is old"* (Genesis 18, 12:13). But when God informs Abraham of Sarah's reaction he quotes her

as saying *“Will I really have a child, now that I am old?”*. God obfuscates the fact that it’s *Abraham’s* age she laughed about.

According to the Talmud, this story teaches the importance of “peace in the family” *(shlom-bayit)*. It’s a resource so important that it’s worth God lying to preserve it. *Shlom-bayit* is hard to quantify, but your behavior is as constrained when you’ve lost your partner’s goodwill as if you were down to your last dollar or minute. When maintaining your slack according to formula, don’t forget to count the uncountable resources too.

Earphones

I find it much easier to untangle my earphones when I hold the entire cable in a loose lump in my hand so that none of the wires are pulled taut.

Exploration

One of the main things that slack gives you is *optionality*, the freedom to change your plans. The value of optionality varies a lot depending on what one is up to, whether you’re exploring or exploiting. “Exploit” is when a single best option is available to you, and you pursue it single-mindedly. For example, I’m starting an internship at a dream company in a few weeks, and I will care about nothing except getting a full-time position there. I won’t be intersted in other employment options, and I won’t need slack for anything besides work.

“Exploration” is when a lot of paths are open, when there’s great potential but little certainty. That’s when slack is valuable, it allows you to pursue the opportunities. I learned to appreciate slack in my own life after messing up a critical exploration phase due to slacklessness, my college years.

When I was 18, I joined a very selective academic officer training program. We pursued a double degree in math and physics condensed into three years, along with intensive military training and enough chores to be a bummer. We had negative slack: no money, no freedom, and a daily to-do list that would take about 20 hours to complete, but only if you were fresh off 8 hours of sleep. I dropped out after a hectic two years. I realized that not only did I remember almost nothing from the classes, I didn’t even know if I liked physics, or the army, or if I actually wanted to be an officer.

The same year my wife-to-be enrolled in a community college without much pressure to do anything other than study and to try stuff. She made lifelong friends, learned Japanese, tried out a bunch of subjects and eventually discovered and fell in love with biology. Then, she could shift fully into exploit mode: she aced all the available biology class, transferred to a good university where completed a degree in biology in two years, and got into a leading graduate program. She became a biologist because she had several years when she didn’t have to decide what she would become.

I made up for my slackless undergrad experience by going to a slack-friendly business school. I had time to play every single intramural sport, go on a lot of drunk dates, and become a regular writer for a satire magazine which I spent more time on than all my homework combined. Writing satire later turned into a paying gig, a short but exciting stand-up career in NYC, and eventually Putanumonit when I couldn’t get anyone to pay to hear my jokes. This blog only exists because in the last few years I’ve guarded my slack jealously.

And yet, I see smart people in elite universities fall into the same trap I did originally. Very prestigious schools are very competitive, and competition will incinerate every bit of slack you have. Heavy course loads leave students little slack to fool around with satire, squash, or even fooling around. Once you start pursuing a major there’s little slack to learn anything else, and once you graduate with a load of debt there’s no slack to do anything but take the first paying job on offer. A less prestigious university that requires half the time, the effort

and the money of an elite school often offers a better education simply by leaving you with slack and the freedom to explore.

The same is true for other exploration activities like travel (on long trips, I try to leave 50% of the days unscheduled), job hunting, and dating. Slacklessness brings desperation, and desperation leads to making the sort of choices that your friends will shake their heads about a decade later. Fight for your slack, and give yourself it.

Amos Tversky Said

The secret to doing good research is always to be a little underemployed. You waste years by not being able to waste hours.

How much time did he waste coming up with this pithy aphorism? It was worth it.

The Mad Scientist Decision Problem

Consider Alice, the mad computer scientist. Alice has just solved general artifical intelligence and the alignment problem. On her computer she has two files, each containing a seed for a superintelligent AI, one of them is aligned with human values, the other one is a paperclip maximizer. The two AIs only differ in their goals/values, the rest of the algorithms, including decision procedures, are identical.

Alice decides to flipp a coin. If the coin comes up heads, she starts the friendly AI, and if it comes up tails, she starts the paperclip maximizer.

The coin comes up heads. Alice starts the friendly AI, and everyone rejoice. Some years later the friendly AI learns about the coinflip and of the paperclip maximizer.

Should the friendly AI counterfactually cooperate with the paperclip maximizer?

What does various decision theories say in this situation?

What do you think is the correct answer?